Stability concepts and their applications

Imre Fekete

Ph.D. Thesis

Eötvös Loránd University

2015

Eötvös Loránd University

Doctoral School of Mathematics
Applied Mathematics Program

Ph.D. Thesis

Stability concepts and their applications

Author:
Imre Fekete
Ph.D. Candidate

Supervisor:
Prof. István Faragó
DHAS

School Leader: Prof. Miklós Laczkovich, MHAS
Program Leader: Prof. György Michaletzky, DHAS

Department of Applied Analysis
and Computational Mathematics

Budapest
2015

"Mathematics is the most beautiful and most powerful creation of the human spirit."

Stefan Banach

Table of Contents

List of Figures

List of Tables

Introduction

This dissertation deals with stability concepts for operator equations and their possible application areas in theoretical numerical analysis. This thesis is based on the Author's papers [27], [24], [29], [28], the accepted paper [19] and the preprint [25]. The thesis consists of five chapters.

In Chapter 1 we set the problem in an abstract setting and introduce the basic notions in numerical analysis. Furthermore, we show what is the relation between consistency and convergence for nonlinear operator equations.

In Chapter 2 we deal with N-stability notion and we show its possible application areas in theoretical numerical analysis. In Section 2.2 it turns out that linear multistep methods and the zero-stability notion fits into our framework and we regain the classical results from the literature. In Section 2.3 we offer a new and effective tool in order to verify stability results for time-dependent problems. The benchmark problems are reaction-diffusion and transport problems. In Section 2.4 we consider nonlinear evolution equations whose solution is given by a nonlinear semigroup. We show that the definition of nonlinear semigroups already contains a sort of time discretization, the implicit Euler method, which leads to N-stable discrete problems when applied together with certain convergent space discretizations. Moreover, we propose a more general time discretization, being the nonlinear counterpart of the rational approximations in the linear case and show its N-stability as well.

In Chapter 3 we deal with other stability notions. First, in Section 3.1 we give an example to motivate local type stability notions. In Section 3.2 we show the benefits of this notion in theory as well as from the application point of view. In Section 3.3 we prove theoretical results for Trenogin's stability notion and we improve his results. In the end of this chapter we give some comments on other stability notions.

In the fist part of Chapter 4 we extend the previously given pointwise (local) definitions to the set (global) ones. Under reasonable assumptions we prove the set version of the basic theorem of numerical analysis. In the second part we show the relation between the basic notions. Based on the previous results of this section we can theoretically answer the most important cases and we can also give examples in the Appendix Section A.3.

In Chapter 5 we precisely summarize our results for each chapter.

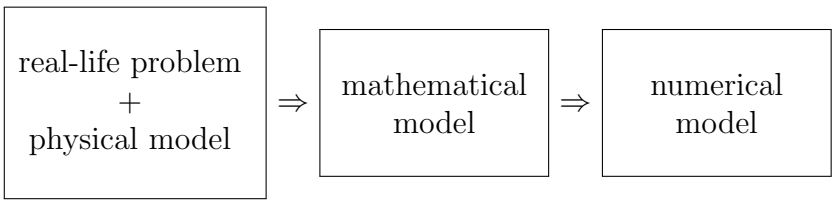
CHAPTER 1

Basic notions in numerical analysis

Many phenomena in nature can be described by principle based mathematical models which consist of functions of a certain number of independent variables and parameters. In particular, these models often consist of equations, usually containing a large variety of derivatives with respect to the variables. Typically, we are not able to give the solution of the mathematical model in a closed (analytical) form, therefore we construct some numerical and computer models that are useful for practical purposes.

The ever-increasing advances in computer technology have enabled us to apply numerical methods to simulate plenty of physical phenomena in science and engineering. As a result, numerical methods do not usually give the exact solution to the given problem, they can merely provide approximations, getting closer and closer to the solution with each computational step. Numerical methods are generally useful only when they are implemented on a computer via a computer programming language. This way, with detailed and realistic mathematical models and numerical methods, it is possible to gain quantitative (and also qualitative) information for a multitude of phenomena and processes in physics and technology. The application of computers and numerical methods has become ubiquitous. Computations are often cheaper than experiments; experiments can be expensive, dangerous or downright impossible. Real-life experiments can often be performed on a small scale only and that makes their results less reliable.

The above described modelling process of real-life phenomena can be illustrated as follows:



This means that the complete modelling process consists of three steps. This dissertation analyses the step when we transform the mathematical (usually continuous) model into numerical (usually discrete) models and it also investigates the numerical models.

The discrete model usually yields a sequence of discrete tasks. During the construction of numerical models the basic requirements are the following:

- ◇ Each discrete problem in the numerical model is a well-posed problem, i.e.
  - there exists a sequence of solution (existence),
  - the solution is unique (uniqueness),
  - the solution depends continuously on the data (stability).
- ◇ In the numerical model we can efficiently compute the numerical solution.
- ◇ The sequence of the numerical solutions is convergent.
- ◇ The limit of this sequence is the solution of the original problem.

Our aim is to guarantee that this step does not cause any significant loss of the information.

1.1 Setting the problem

Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed spaces and $F : \mathrm{dom}(F) \subset X \to Y$ be a (possibly unbounded and nonlinear) operator. When we model some real-life phenomenon with a mathematical model, we end up investigating the problem

$$F(u) = 0 \quad \text{for} \quad u \in \mathrm{dom}(F). \tag{1.1}$$

The abstract framework of investigating this kind of equations was first introduced by Stetter in [56] and Trenogin in [62]. Later Sanz-Serna, Palencia and López-Marcos systematically studied the modified version of Stetter's framework [44, 45, 50, 51, 53, 54]. Another possible treatment can be found in [27]. The framework and definitions of this section are based on the latter one.

Definition 1.1.1. *Problem (1.1) can be given as a triplet $\mathscr{P} = (X, Y, F)$. We will refer to it as problem $\mathscr{P}$.*

Example 1.1.1. Consider the following initial value problem:

$$u'(t) = f(u(t)), \quad t \in (0, T], \tag{1.2}$$

$$u(0) = u_0, \quad u_0 \in \mathbb{R}, \tag{1.3}$$

where $f \in C(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function. Then problem $\mathscr{P}$ can be given as

$$X = C^1([0,T]), \quad \|u\|_X = \max_{t\in[0,T]} |u(t)|,$$

$$Y = C([0,T]) \times \mathbb{R}, \quad \left\| \begin{pmatrix} u \\ u_0 \end{pmatrix} \right\|_Y = \max_{t\in(0,T]} (|u(t)|) + |u_0|,$$

$$F(u) = \begin{pmatrix} u'(t) - f(u(t)) \\ u(0) - u_0 \end{pmatrix}. \qquad \clubsuit$$

In the sequel we assume that that there exists a unique solution of (1.1). It will be denoted by u^*. However, in case of concrete applied problems *we must prove* the existence of $u^* \in \mathrm{dom}(F)$. Generally the proof is not constructive, see e.g. [40]. Even if it is possible to solve directly, the realization of the solving process is very difficult or even impossible.

Luckily, we only need a good approximation for the solution of problem (1.1), since our model is already a simplification of the real-life phenomenon. So our ultimate goal is to replace problem (1.1) with a sequence of simpler problems. In order to achieve this goal, it is enough to use some discretization and numerical methods. The basic requirements of these were formulated in the earlier part of this section.

The sequence of simpler problems mathematically means nothing else but defining an index set $\mathbb{I} \subset \mathbb{N}^p$ for $p \in \mathbb{N}$, normed spaces $(X_n, \|\cdot\|_{X_n})$, $(Y_n, \|\cdot\|_{Y_n})$ and sequence of operators $F_n : \mathrm{dom}(F_n) \subset X_n \to Y_n$. Then one can consider the sequence of problems

$$F_n(u_n) = 0 \quad \text{for} \quad u_n \in \mathrm{dom}(F_n) \quad \text{and} \quad n \in \mathbb{I}. \tag{1.4}$$

Definition 1.1.2. *The sequence $\mathscr{N} = (X_n, Y_n, F_n)_{n\in\mathbb{I}}$ is called a numerical method if it generates a sequence of problems (1.4).*

If there exists a unique solution of (1.4), then it will be denoted by u_n^*.

Example 1.1.2. Continuing Example 1.1.1 we can define the numerical method $\mathscr{N}$ as

$$X_n = \mathbb{R}^{K+1}, \quad v_n = (v_0, v_1, \dots, v_K) \in X_n : \|v_n\|_{X_n} = \max_{k=0,\dots,K} |v_k|,$$

$$Y_n = \mathbb{R}^{K+1}, \quad y_n = (y_0, y_1, \dots, y_K) \in Y_n : \|y_n\|_{Y_n} = |y_0| + \max_{k=1,\dots,K} |y_k|,$$

$F_n : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$ and for any $v_n = (v_0, v_1, \dots, v_K) \in \mathbb{R}^{K+1}$ it acts as

$$[F_n(v_n)]_k = \begin{cases} \dfrac{K}{T}(v_k - v_{k-1}) - f(v_{k-1}), & k = 1, \dots, K, \\ \\ v_0 - u_0. & k = 0. \end{cases}$$

♣

Remark 1.1.1. One of our goals is to give an estimation to the element $u^* - u_n^*$, since this subtraction represents the error. It is easy to see that in spite of the introduced definitions we have the following difficulties:

(a) Comparison of u^* and u_n^*, since these might be found in different spaces.

(b) Comparison seems to be impossible, since u^* is not known.

In order to treat Remark 1.1.1 (a) and make connection between the problems (1.1) and (1.4) we give the following definition.

Definition 1.1.3. *Let there be the mappings $\varphi_n : X \to X_n$ and $\psi_n : Y \to Y_n$ for all $n \in \mathbb{I}$. Then the sequence $\mathscr{D} = (\varphi_n, \psi_n, \Phi_n)_{n\in\mathbb{I}}$ is called a discretization, where*

$$\Phi_n : \{F : \mathrm{dom}(F) \to Y \mid \mathrm{dom}(F) \subset X\} \to \{F_n : \mathrm{dom}(F_n) \to Y_n \mid \mathrm{dom}(F_n) \subset X_n\}.$$

Assumption 1.1.1.

(a) For the mapping ψ_n the relation $\psi_n(0) = 0$ holds.

(b) $\dim(X_n) = \dim(Y_n) < \infty$.

Remark 1.1.2. Obviously, when ψ_n are linear operators, then Assumption 1.1.1 (a) is automatically satisfied. Assumption 1.1.1 (b) is important because of the application point of view and the well-posedness of problem (1.4).

Example 1.1.3. Define the equidistant grid

$$\{t_k = k\tau, \text{ where } k = 0, \dots, K \text{ and } \tau = T/K\}.$$

on the interval $[0, T]$. Based on Examples 1.1.1 and 1.1.2, in Definition 1.1.3 we define discretization $\mathscr{D}$ as

$$\varphi_n(y) : C^1([0,T]) \to \mathbb{R}^{K+1} \text{ such that } [\varphi_n(y)]_k = y(t_k), \quad k = 0, 1, \dots, K,$$

$$\psi_n(y) : C([0,T]) \times \mathbb{R} \to \mathbb{R}^{K+1} \text{ such that}$$

$$[\psi_n(y)]_k = \begin{cases} y\left(t_{k-1}\right), & k = 1, \dots, K, \\ y(t_0), & k = 0. \end{cases}$$

In order to give Φ_n, we define the mapping $\Phi_n : C^1([0,T]) \to \mathbb{R}^{K+1}$ in the following way:

$$\left[\left(\Phi_n(F)\right)\left(\varphi_n(u)\right)\right]_k = \begin{cases} \dfrac{u(t_k) - u(t_{k-1})}{\tau} - f(u(t_{k-1})), & k = 1, \dots, K, \\ u(t_0) - u_0, & k = 0. \end{cases}$$

Thus, we fully discretized in this abstract framework the equations (1.2)-(1.3). ♣

Remark 1.1.3. In the sequel we will not determine exactly the operator Φ_n in Definition 1.1.3. It will be a matter of course.

To overcome the difficulty mentioned in Remark 1.1.1 (b), the usual idea is to introduce the notions of consistency and stability, which are controllable. The notion of stability is independent of the solution of the original problem (1.1). From the linear literature it is known that generally convergence can be replaced with these two notions. Sometimes this popular "recipe" is summarized in the implication

$$\text{Consistency} + \text{Stability} \Rightarrow \text{Convergence}. \tag{1.5}$$

This implication is also known in the literature as the "basic theorem of numerical analysis". Motivated by the linear case we would like to introduce and investigate these notions in an abstract framework and we try to shed some light on implication (1.5) in the nonlinear case, too.

In this case the naturally arising questions are the following:

◇ How shall we define consistency and stability to ensure implication (1.5)?

◇ Are consistency and/or stability necessary for convergence?

In sense of Definition 1.1.1, 1.1.2 and 1.1.3 we can imagine the numerical process as in Figure 1.1. A similar figure can be found in [56] and [31].

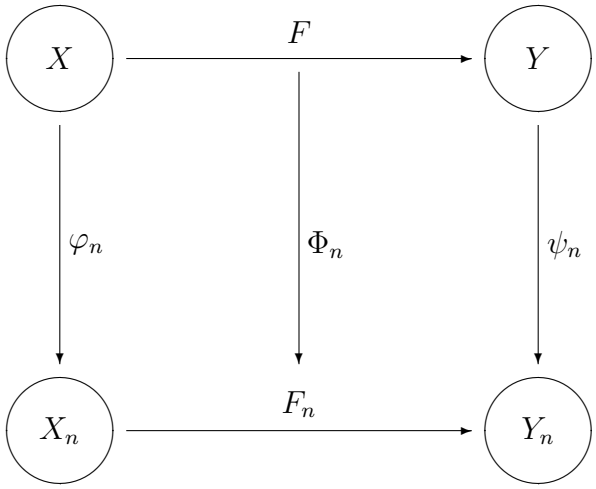


Figure 1.1. *The general scheme of numerical process.*

1.2 Basic definitions

In this section we give the definitions of convergence and consistency and show the connection between them. This leads to the motivation of the stability notion.

1.2.1 Convergence

We would like to compare the solutions of (1.1) and (1.4). Since these elements belong to different spaces we use the mapping $\varphi_n : X \to X_n$ in order to measure the distance between them in X_n.

Definition 1.2.1. *The element $e_n = \varphi_n(u^*) - u_n^* \in X_n$ is called global discretization error.*

Our goal is to guarantee arbitrary smallness of the global discretization which can be generally achieved by increasing n. It motivates the following definition.

Definition 1.2.2. *The discretization $\mathscr{D}$ applied to problem $\mathscr{P}$ is called convergent if*

$$\lim_{n\to\infty} \|e_n\|_{X_n} = 0 \tag{1.6}$$

holds. When

$$\|e_n\|_{X_n} = \mathcal{O}(n^{-p})$$

we say that the order of the convergence is p.

Remark 1.2.1. Definition 1.2.2 depends on the approximation capabilities of the space sequence $(X_n)_{n\in\mathbb{I}}$. Therefore, in this case the so called norm consistency assumption for any arbitrary chosen $f \in X$

$$\lim_{n\to\infty} \|\varphi_n(f)\|_{X_n} = \|f\|_X \tag{1.7}$$

is logical.

Remark 1.2.2. There is another possibility to compare the solutions. The normed space X might be more natural at first sight. Using the mapping $\bar{\varphi}_n : X_n \to X$ we are able to define the notion of convergence as

$$\lim_{n\to\infty} \|u^* - \bar{\varphi}_n(u_n^*)\|_X = 0. \tag{1.8}$$

A similar condition to (1.7) can be given in this case. Namely, it is the condition that $\lim_{n\to\infty} \bar{\varphi}_n(\varphi_n(f)) = f$ for all $f \in X$. Then the whole process can be imagined as in Figure 1.2. The difficulty of this approach is that the convergence depends on the numerical method and on the mappings $\bar{\varphi}_n$. Therefore, as most of the authors, we choose the earlier defined notion.

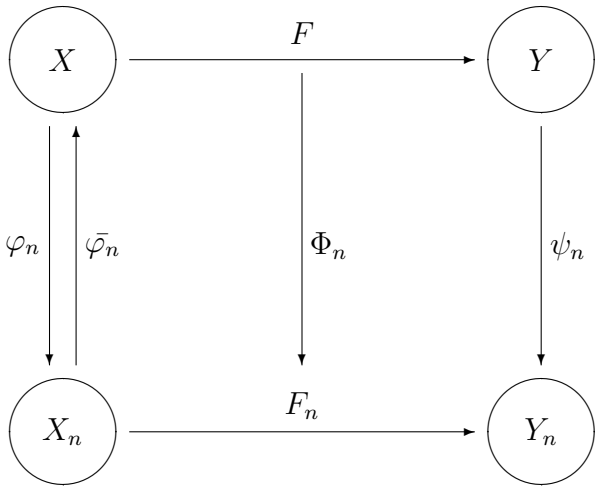


Figure 1.2. *The general scheme of numerical process in case of mapping $\bar{\varphi}_n$.*

1.2.2 Consistency

Independently of the form of the definition of the global error it is hardly applicable in practice, since the knowledge of the exact solutions are assumed. Hence, we introduce the notion of consistency which may help us in getting information about the behaviour of the global discretization error.

Definition 1.2.3. *The element $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in Y_n$ is called local discretization error on the element v.*

Remark 1.2.3. A special role is played by the behaviour of $l_n(v)$ on the solution of the problem (1.1). Using Assumption 1.1.1 (a) we get for solution of (1.1) that $l_n(u^*) = F_n(\varphi_n(u^*)) - \psi_n(F(u^*)) = F_n(\varphi_n(u^*))$. For simplicity we will use the notation l_n for $l_n(u^*)$.

Definition 1.2.4. *The discretization $\mathscr{D}$ applied to problem $\mathscr{P}$ is called consistent on the element $v \in \mathrm{dom}(F)$ if*

i, $\varphi_n(v) \in \mathrm{dom}(F_n)$ *holds from some index,*

ii, the relation

$$\lim_{n\to\infty} \|l_n(v)\|_{Y_n} = 0 \tag{1.9}$$

holds.

If

$$\|l_n(v)\|_{X_n} = \mathcal{O}(n^{-p}),$$

then we say that the order of the consistency on the element v is p.

Remark 1.2.4. In the sequel, the consistency on u^* and its order will be called consistency and order of consistency.

Fix some element $v \in \mathrm{dom}(F)$. Then we can transform it into the space Y_n in two different ways (c.f. Figure 1.1). The magnitude $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in Y_n$ in (1.9) plays an important role in numerical analysis, since it characterizes the difference of these two directions for the element v. Hence, the consistency on the element v yields that in limit the diagram of Figure 1.1 is commutative.

Remark 1.2.5. One might ask whether consistency implies convergence. Example A.1.1 shows that this is not true in general. Thus, convergence cannot be replaced by consistency in general. Assuming the existence of the inverse operator F_n^{-1} we can easily get the relation

$$e_n = \varphi_n(u^*) - u_n^* = F_n^{-1}(F_n(\varphi_n(u^*))) - F_n^{-1}(0) = F_n^{-1}(l_n(u^*)) - F_n^{-1}(0).$$

It shows the connection between the global and local discretization errors. This relation suggests that the consistency (i.e., the convergence to of the local discretization error l_n to zero) can provide the convergence (i.e., the approach of e_n to zero) when $(F_n^{-1})_{n\in\mathbb{I}}$ has good behaviour. Such a property is the Lipschitz continuity: it would be useful to assume that the functions F_n^{-1} uniformly satisfy the Lipschitz condition at the point $0 \in Y_n$. However, generally at this point we have no guarantee even to the existence of F_n^{-1}, thus we provide this with some property of the functions F_n, without assuming their invertibility.

CHAPTER 2

N-stability and its applications

Convergence yields that the global discretization error e_n tends to 0. Having consistency, we have information about the local discretization error only. Intuitively, this means that when $l_n(u^*)$ is small, then e_n should be small, too. Since u^* is unknown, in first approach we require this property for any pairs in $\mathrm{dom}(F_n)$. This demand implies the requirement

$$\|z_n - w_n\|_{X_n} \leq C(n)\|F_n(z_n) - F_n(w_n)\|_{Y_n} \tag{2.1}$$

holds for arbitrary $z_n, w_n \in \mathrm{dom}(F_n)$.

The problem with this approach is that the constant $C(n)$ in (2.1) can grow into infinity as n tends to ∞. In order to guarantee the well-posedness of the discrete problems it means that the constant in (2.1) has to be uniformly bounded.

Therefore, we consider the estimate

$$\|z_n - w_n\|_{X_n} \leq C\|F_n(z_n) - F_n(w_n)\|_{Y_n} \tag{2.2}$$

holds for arbitrary $z_n, w_n \in \mathrm{dom}(F_n)$ and the constant C is independent of the mesh size parameter. This idea leads to make the first attempt to define the nonlinear stability notion.

Definition 2.0.5. *The discretization $\mathscr{D}$ is called N-stable on problem $\mathscr{P}$ if there exists a positive stability constant C such that for each $z_n, w_n \in \mathrm{dom}(F_n)$ the estimate (2.2) holds.*

Definition 2.0.5 originally defined by López-Marcos and Sanz-Serna in [44]. In the sequel we will refer to this notion as the natural stability (N-stability) for the nonlinear case.

For nonlinear problems the following result is true.

Theorem 2.0.1. *We assume that*

i, there exists the solution of problems (1.1) and (1.4),

ii, discretization $\mathscr{D}$ is consistent in order p on element u^* and N-stable with the stability constant C,

iii, for the mapping ψ_n the relation $\|\psi_n(0)\|_{Y_n} = \mathcal{O}(n^{-p})$ holds.

Then discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$ and the order of convergence is not less than the order of consistency.

Proof. Using i, and Definition 2.0.5 we have the estimation

$$\begin{aligned}\|e_n\| &\leq C\|F_n(\varphi_n(u^*)) - F_n(u_n^*)\|_{Y_n} \\ &\leq C\|F_n(\varphi_n(u^*)) - \psi_n(F(u^*))\|_{Y_n} + C\|\psi_n(F(u^*)) - F_n(u_n^*)\|_{Y_n},\end{aligned}$$

where the first term converges to zero as n goes to infinity due to consistency and the second term converges to zero because of i, and iii,. Hence, the order of convergence is not less than the order of consistency. ■

Remark 2.0.6. The assumption iii, of Theorem 2.0.1 is weaker than Assumption 1.1.1 (b).

This result shows the role of both stability and consistency for obtaining convergence in case of nonlinear operator equations.

2.1 Linear stability as a special case

The relationship between stability and convergence for linear problems hinted by Courant, Friedrichs and Lewy in the 1920's [16], identified more clearly by von Neumann [15] in the 1940's and brought into organized form by Lax and Richtmyer in the 1950's as the Lax (or sometimes Lax–Richtmyer–Kantorovich [42]) equivalence theorem. From the formulation of the main theorem it turns out that these two directly checkable conditions (i.e., consistency and stability) serve together convergence.

First of all we consider the sequence of linear problems

$$L_n u_n = 0, \quad \text{for } u_n \in \operatorname{dom}(L_n), \tag{2.3}$$

where for each $n \in \mathbb{I}$ the operators $L_n : \operatorname{dom}(L_n) \to Y_n$ are linear. Naturally, we always assume the solvability of the problems (2.3), i.e. the existence of the operators $L_n^{-1} : Y_n \to \operatorname{dom}(L_n)$.

Definition 2.1.1. *The discretization $\mathscr{D}$ is called stable on the linear problem $\mathscr{P}$ if there exists a positive stability constant C such that for each $s_n \in \operatorname{dom}(L_n)$*

$$\|s_n\|_{X_n} \leq C\|L_n s_n\|_{Y_n} \tag{2.4}$$

holds.

It is easy to see that Definition 2.1.1 is the special case of Definition 2.0.5. Hence, N-stability can be viewed as the natural extension of Definition 2.1.1.

Remark 2.1.1. The bound (2.4) implies three basic properties:

i, For any problems (2.3) the relation (2.4) shows that $L_n s_n = 0$ implies that $s_n = 0$, i.e., L_n is injective and hence L_n^{-1} exists on the entire space Y_n by Assumption 1.1.1 (b). If L_n is surjective, then the stability bound implies the existence and uniqueness of the solutions of (2.3).

ii, Due to i, and (2.4), we have

$$\|L_n^{-1} s_n\|_{X_n} \leq C \|s_n\|_{Y_n}$$

for all $s_n \in Y_n$. Therefore the uniform norm estimate

$$\|L_n^{-1}\|_{B(Y_n, X_n)} \leq C$$

holds. Sometimes it is referred to as linear stability after Kantorovich [38].

iii, In view of (2.4), we obtain the "basic theorem of numerical analysis". In fact, due to the linearity of L_n we get

$$\|e_n\|_{X_n} = \|\varphi_n(u^*) - u_n^*\|_{X_n} \leq C \|L_n \varphi_n(u^*)\|_{Y_n} = C \|l_n(u^*)\|_{Y_n},$$

where we use Assumption 1.1.1 (a). Obviously for consistent methods in order p this implies the convergence in order p, too.

Hence, the linear stability notion implies some basic results. However, obtaining these consequences we exploit the linearity of the operators L_n.

Remark 2.1.1 (i) and (ii) show that the linear stability notion is implied by N-stability. On the other hand, the reverse implication is also true, since

$$\|s_n\|_{X_n} = \|L_n^{-1} L_n s_n\|_{Y_n} \leq \|L_n^{-1}\|_{B(Y_n, X_n)} \|L_n s_n\|_{Y_n} \leq C \|L_n s_n\|_{Y_n}.$$

Thanks to these results we can state that for linear problems N-stability is equivalent to the linear stability notion.

2.2 Operator Form of Multistep Methods

Let us consider the initial-value problem

$$u'(t) = f(t, u(t)), \tag{2.5}$$
$$u(0) = u_0, \tag{2.6}$$

where $f : \Omega \to \mathbb{R}^d$ is a Lipschitz continuous function, $\Omega \subset (0, T] \times \mathbb{R}^d$ and $u_0 \in \mathbb{R}^d$ is the initial-value vector. For the sake of simplicity we will consider the scalar case. The generalization for ODEs is straightforward.

Then, similarly to Exapmle 1.1.1 we can rewrite equations (2.5)-(2.6) in the introduced framework with the following choices:

- $\diamond$ $X = C^1([0, T])$,
- $\diamond$ $Y = C((0, T]) \times \mathbb{R}$,
- $\diamond$ the mapping $L : X \to Y$ on an element $w \in X$ acts as

$$[Lw](t) = \begin{cases} w'(t) - f(t, w(t)), & t \in (0, T], \\ w(0), & t = 0. \end{cases} \tag{2.7}$$

Due to the validity of existence and uniqueness of problem (2.5)-(2.6) the operator (2.7) is injective (see A.2.1). Therefore in case of a given function $g(t)$ the problem $Lu = g$ has a unique solution. Let g be the following choice

$$g(t) = \begin{cases} 0, & t \in (0, T], \\ u_0, & t = 0. \end{cases}$$

The equation $Lu = g$ can be rewritten in the form of (1.1) in case of appropriately restriction of the domain of the linear operator. Namely, we define the mapping L on an element $w \in X$ as

$$[Lw](t) = w'(t) - f(t, w(t)), \ t \in (0, T]$$

with the domain

$$\mathrm{dom}(L) = \{w \in X) \mid w(0) = u_0\}.$$

Thus, the scalar version of (2.5)-(2.6) can be rewritten in the form of (1.1).

Remark 2.2.1. In Example 1.1.1 we gave a different form in order to rewrite (2.5)-(2.6) in the form of (1.1).

2.2.1 Zero-stability of one-step methods

Zero-stability is one of the basic concepts in the numerical theory of ODEs. However, in many cases most of the authors do not give precise definition of zero-stability for linear one-step methods or simply they skip this definition (e.g. [14], [34], [30], [43], [59], [33]). They just intuitively describe us that "zero-stability can be determined by merely considering the method's behaviour when applied to the trivial differential equation $y' = 0$; it is for this reason that the concept of stability is referred to as zero-stability" or "a method is stable if the corresponding difference equation has only bounded solutions".

In this section our main goal is to use the benefits of the previously introduced framework and N-stability in order to prove theoretical results. First, we define the spatial grid as

$$\omega_\tau := \{0 = t_0 < t_1 < \ldots < t_{K-1} < t_K = T\}. \tag{2.8}$$

Furthermore, we introduce the notation

$$\omega_\tau^0 := \omega_\tau \setminus \{0\}.$$

Let us define the mappings φ_n and ψ_n as grid functions. The vector spaces defined on ω_τ and ω_τ^0 grids of the grid functions are denoted by $\mathbb{F}(\omega_\tau)$ and $\mathbb{F}(\omega_\tau^0)$, respectively. Furthermore, let the step-size be defined as $\tau_j = t_{j+1} - t_j, \ j = 0, \ldots, K-1$ and $\tau = T/K$. Suppose that there exists a positive constant c such that for all K the estimate $\tau_j \leq c\tau$ holds for all $j = 0, \ldots, K-1$. We also suppose that the fixed point $t^\star \in (0, T]$ is an element of all grids. On a fixed grid the index k denotes the index for which $\tau_0 + \ldots + \tau_{k-1} = t^\star$.

Let us choose the normed spaces X_n and the operator:

- $X_n = \mathbb{F}(\omega_\tau)$,
- $Y_n = \mathbb{F}(\omega_\tau)$,
- $L_n : X_n \to Y_n$ on an element $w_n \in X_n$ as

$$[L_n w_n](t_k) = \begin{cases} \Phi(\tau_k, t_{k-1}, w_n(t_{k-1}), w_n(t_k)), & t_k \in \omega_\tau^0, \\ w_n(0), & t_k = 0, \end{cases} \tag{2.9}$$

where Φ denotes the given one-step method.

Remark 2.2.2. In order to realize the method we have to assume that the first three variables are fixed and Φ can be invertible in the fourth variable. It means that the function $s \mapsto \Phi(\tau^\star, t^\star, \omega^\star, s)$ is invertible.

Since the operator (2.9) is injective (see A.2.2), in case of a given function $g_n(t)$ the problem $L_n u_n = g_n$ has a unique solution. Let g_n be the following choice

$$g_n(t_k) = \begin{cases} 0, & t_k \in \omega_\tau^0, \\ u_0, & t_k = 0. \end{cases}$$

The equation $L_n u_n = g_n$ can be rewritten in the form of (1.4) in case of appropriate restriction of the domain of the linear operator. Namely, we define the mapping L_n on an element $w_n \in X_n$ as

$$[L_n w_n](t_k) = \Phi(\tau_k, t_{k-1}, w_n(t_{k-1}), w_n(t_k)), \ t_k \in \omega_\tau^0. \tag{2.10}$$

Remark 2.2.3. Since the defined operator (2.10) maps from $\mathbb{F}(\omega_\tau)$ to $\mathbb{F}(\omega_\tau^0)$ and $\dim(\mathbb{F}(\omega_\tau)) \neq \dim(\mathbb{F}(\omega_\tau^0))$, it is not injective on $\mathbb{F}(\omega_\tau)$. Consequently there does not exist a unique solution of problem (1.4).

Due to the previous observation we restrict the domain of operator L_n such that $\mathrm{dom}(L_n) \subset X_n$ and $\dim(\mathrm{dom}(L_n)) = \dim(\mathbb{F}(\omega_\tau^0))$. The required domain is

$$\mathrm{dom}(L_n) := \{w_n \in X_n \mid w_n(t_0) = u_0\}. \tag{2.11}$$

It follows that using (2.10) and (2.11) we can rewrite one-step methods in the form of (1.4). Now we would like to give an appropriate zero-stability definition.

Definition 2.2.1. *The operator (2.10) is called zero-stable (0-stable) if there exist positive constants τ_0 and C such that for all $\tau < \tau_0$ and for arbitrary grid functions $z_n, w_n \in \mathrm{dom}(L_n)$ the estimation*

$$||z_n - w_n||_\infty \leq C\{|z_n(t_0) - w_n(t_0)| + \max_{1 \leq k \leq K} |[L_n z_n](t_k) - [L_n w_n](t_k)|\} \tag{2.12}$$

holds.

Theorem 2.2.1. *The zero-stable operator (2.10) is invertible on domain*

$$\mathrm{dom}(L_n) := \{w_n \in X_n \mid w_n(t_0) \text{ fixed}\}. \tag{2.13}$$

Proof. We have to show that the operator (2.10) is injective on domain (2.13). This follows from the definition of zero-stability. In case of $L_n z_n = L_n w_n$ we have

$$\max_{1\le k\le K} \left|[L_n z_n](t_k) - [L_n z_n](t_k)\right| = 0.$$

On the other hand, since $z_n, w_n \in \mathrm{dom}(L_n)$, $z_n(t_0) - w_n(t_0) = 0$. Taking into account (2.12) we have $||z_n - w_n||_\infty = 0$, i.e. $z_n = w_n$. ■

Theorem 2.2.2. *Assume that*

i, there exists the solution of problem (1.1),

ii, discretization $\mathscr{D}$ *is consistent in order* p *(described by operator* (2.10) *with domain* (2.11) *) and zero-stable.*

Then discretization $\mathscr{D}$ *is convergent on problem* $\mathscr{P}$ *and the order of convergence is not less than the order of consistency.*

Proof. In order to prove this theorem we would like to use Theorem 2.0.1. Since the sequence of operator equations (1.4) has a unique solution on domain (2.11), the first assumption of Theorem 2.0.1 is fulfilled. As a second step we show that estimate (2.12) means that operator (2.10) is N-stable, too. To prove this property we have to choose appropriately the normed spaces and the corresponding norms. We summarize this in Table 2.1.

$\varphi_n,\ \psi_n$	Grid functions
$\mathrm{dom}(L_n)$	(2.11)
X_n	$\mathbb{F}(\omega_\tau)$
Y_n	$\mathbb{F}(\omega_\tau^0)$
$\|v_n\|_{X_n}$	$\max\limits_{1\le k\le K} \|v_n(t_k)\|$
$\|v_n\|_{Y_n}$	$\|v_n(t_0)\| + \max\limits_{1\le k\le K} \left\|[L_n v_n](t_k) - [L_n v_n](t_k)\right\|$

Table 2.1. *How to choose operators, normed spaces and corresponding norms to prove N-stability in case of one-step methods.*

So the assumptions of Theorem 2.0.1 are fulfilled which proves the statement of this theorem. ■

There are precise zero-stability definitions in the literature. In the following table we summarize how these definitions are related to N-stability and fit into our framework. Since the grid functions φ_n, ψ_n and the normed spaces X_n, Y_n are the same, there we give only the corresponding norm.

Zero-stability	$\|\|v_n\|\|_{X_n}$	$\|\|v_n\|\|_{Y_n}$
Gautschi [30]	$\max_{0\le k\le K} \|v_n(t_k)\|$	$\|v_n(t_0)\| + \max_{1\le k\le K} \|[L_n v_n](t_k)\|$
Süli [59]	$\max_{0\le k\le K} \|v_n(t_k)\|$	$\|v_n(t_0)\|$

Table 2.2. *Classical one-step zero-stablity notions in our framework.*

2.2.2 Zero-stability of multistep methods

We would like to apply a similar operator approach in order to write s-step linear multistep methods in a general form and show their 0-stability. Furthermore, we would like to make the connection between these notions and Dahlquist's classical stability results [64].

Using the notations of Section 2.2.1 let us choose the normed spaces in the following way:

- $X_n = \mathbb{F}(\omega_\tau)$,
- $Y_n = \mathbb{F}(\omega_\tau)$.

Taking into account the observation in Remark 2.2.3, in this case we define the mapping L_n on an element $w_n \in X_n$ as

$$[L_n w_n](t_k) = \frac{1}{\tau_k} \sum_{j=0}^{s} \alpha_j w_n(t_{k-j}) - \sum_{j=0}^{s} \beta_j f_{k-j}, \ t_k \in \omega_\tau^0 \tag{2.14}$$

with the domain

$$\mathrm{dom}(L_n) := \{w_n \in X_n \mid w_n(t_l) \text{ fixed for all } \ l = 0, 1, \ldots, s-1\}. \tag{2.15}$$

Then we can define the multistep version of Definition 2.2.1.

Definition 2.2.2. *The operator* (2.14) *is called zero-stable (0-stable) if there exist positive constants* τ_0 *and* C *such that for all* $\tau < \tau_0$ *and for arbitrary grid functions* $z_n, w_n \in \mathrm{dom}(L_n)$ *the estimation*

$$||z_n - w_n||_\infty \le C \left\{ \max_{0\le k\le s-1} |z_n(t_k) - w_n(t_k)| + \max_{s\le k\le K} |[L_n z_n](t_k) - [L_n w_n](t_k)| \right\} \tag{2.16}$$

holds.

As we can see with the choice of $s = 1$ we regain Definition (2.2.1).

Theorem 2.2.3. *The zero-stable operator* (2.14) *is invertible on domain*

$$\mathrm{dom}(L_n) := \{w_n \in X_n \mid w_n(t_l) \text{ fixed for all } l = 0, 1, \ldots, s-1\}. \tag{2.17}$$

Proof. We have to show the operator (2.14) is injective on domain (2.17). This automatically follows from the definition of zero-stability. In case of $L_n z_n = L_n w_n$ we have

$$\max_{s \le k \le K} \left| [L_n z_n](t_k) - [L_n z_n](t_k) \right| = 0.$$

On the other hand, since $z_n, w_n \in \mathrm{dom}(L_n)$, therefore $z_n(t_k) - w_n(t_k) = 0$ for all $k = 0, 1, \ldots, s-1$. Taking into account (2.16) we have $||z_n - w_n||_\infty = 0$, i.e. $z_n = w_n$. ∎

Theorem 2.2.4. *Assume that*

i, there exists the solution of problem (1.1),

ii, the $s-1$ starting values are approximated in order p,

iii, discretization $\mathscr{D}$ is consistent in order p (described by operator (2.14) with domain (2.15)) and zero-stable.

Then discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$ and the order of convergence is not less than the order of consistency.

Proof. In order to prove this theorem we would like to use Theorem 2.0.1. Since the sequence of operator equations (1.4) has a unique solution on domain (2.15), the first assumption of Theorem 2.0.1 is fulfilled. As a second step we show that estimate (2.16) means that operator (2.14) is N-stable, too. To prove this property we have to choose appropriately the normed spaces and the corresponding norms. We summarize this in Table 2.3.

$\varphi_n,\ \psi_n$	Grid functions
$\mathrm{dom}(L_n)$	(2.15)
X_n	$\mathbb{F}(\omega_\tau)$
Y_n	$\mathbb{F}(\omega_\tau^0)$
$\|v_n\|_{X_n}$	$\max\limits_{1 \le k \le K} \|v_n(t_k)\|$
$\|v_n\|_{Y_n}$	$\max\limits_{0 \le k \le s-1} \|v_n(t_k)\| + \max\limits_{s \le k \le K} \left\|[L_n v_n](t_k)\right\|$

Table 2.3. *How to choose operators, normed spaces and corresponding norms to prove N-stability in case of s-step multistep methods.*

So the assumptions of Theorem 2.0.1 are fulfilled which proves the statement of this theorem. ∎

There are precise zero-stability definitions in the literature. In the following table we summarize how these definitions are related to N-stability and fit into our framework. Since the grid functions φ_n, ψ_n and the normed spaces X_n, Y_n are the same, there we give only the corresponding norm.

Zero-stability	$\|\|v_n\|\|_{X_n}$	$\|\|v_n\|\|_{Y_n}$
Gautschi [30]	$\max_{0\le k\le K} \|v_n(t_k)\|$	$\max_{0\le k\le s-1} \|v_n(t_k)\| + \max_{s\le k\le K} \|[L_n v_n](t_k)\|$
Süli [59]	$\max_{0\le k\le K} \|v_n(t_k)\|$	$\max_{0\le k\le s-1} \|v_n(t_k)\|$

Table 2.4. *Classical multistep zero-stablity notions in our framework.*

Remark 2.2.4. From the literature Dahlquist's classical result tells us that a multistep method is stable if and only if its characteristic polynomial satisfies the root condition. Gautschi and Süli proved this statement for Definition 2.2.2 in Theorem 6.3.3. [30] and for the second definition in Table 2.4 in Theorem 12.4. [59], respectively.

2.3 Time-dependent problems

As we mentioned earlier a lot of physical, biological or chemical processes can be fit in this abstract framework (e.g. [1], [2], [49]). In this section we are dealing with two classical problems: reaction-diffusion problems and advection problems. Considering these problems our goal is to show one of the advantages of the N-stability notion. Namely, it can serve as an effective tool for verifying stability properties for time dependent problems.

2.3.1 Reaction-diffusion problems

In chemistry one of the most investigated problems is the reaction-diffusion problem. Reaction-diffusion is a process in which two or more chemicals diffuse over a surface and react with one another to produce stable patterns.

Classical stability results are verified for periodic initial-value reaction-diffusion problems in case of globally Lipschitz continuous forcing function f in several works, e.g. in Ascher [3], Strikwerda [58], [61] and Thomas [60]. Regarding the stability proof, these books use the fact that we know the eigenvalues of the standard matrix replacement of the second derivative operator with periodic boundary conditions. Basic techniques are also introduced e.g., discrete time Fourier transform and von Neumann analysis [58], [60]. The von Neumann approach can be successfully applied in the constant coefficient linear case with periodic boundary conditions or to the Cauchy problem.

Diffusion problem

Consider the following periodic initial-value diffusion problem in one dimension:

$$\partial_t u(t,x) = \partial_{xx} u(t,x), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.18}$$

$$u(t,x) = u(t,x+1), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.19}$$

$$u(0,x) = u^0(x), \qquad x \in \mathbb{R}, \tag{2.20}$$

where $T \in \mathbb{R}^+$. The condition (2.19) yields periodic boundary conditions. Condition (2.20) is the initial-value condition, where u^0 is a given one-periodic function. It is easy to see that the continuous problem (2.18)-(2.20) can be rewritten in the form of (1.1). As we have already mentioned, we assume the existence of the unique, sufficiently smooth solution of the problem (2.18)-(2.20).

Remark 2.3.1. Since the solution is periodic, it is sufficient to determine the solution in one period only.

To create the discretization $\mathscr{D}$ on the above mentioned problem we define both the spatial and time grids, as follows. The spatial grid points are

$$\{x_m = mh, \text{ where } m = 1, \dots, M, \ h = 1/M \text{ and } M \in \mathbb{N}, \ M \geq 2\}$$

and the time levels are

$$\{t_k = k\tau, \text{ where } k = 0, \dots, K \text{ and } \tau = T/K\}.$$

Let us apply an IMEX-type method to (2.18)-(2.20) and we will refer to this method as θ-method.

Remark 2.3.2. We split the the diffusion operator as

$$\partial_{xx} u(t,x) = (1-\theta)\partial_{xx} u(t,x) + \theta \partial_{xx} u(t,x).$$

In this context IMEX-methods mean that the first and second terms are treated explicitly and implicitly, respectively. This technique has a broad literature. The most fundamental references are [5] and [4].

Applying the θ-method to (2.18)-(2.20) for $\theta \in [0,1]$, we gain

$$\frac{u_m^{k+1} - u_m^k}{\tau} - (1-\theta)\frac{u_{m-1}^k - 2u_m^k + u_{m+1}^k}{h^2} - \theta \frac{u_{m-1}^{k+1} - 2u_m^{k+1} + u_{m+1}^{k+1}}{h^2} = 0, \quad (2.21)$$

where $m = 1, \dots, M$, $k = 0, \dots, K-1$ and using the periodic boundary conditions it is obvious that $u_0^k = u_M^k$, $u_1^k = u_{M+1}^k$, $u_0^{k+1} = u_M^{k+1}$ and $u_1^{k+1} = u_{M+1}^{k+1}$. The initial-value condition can be written as

$$u_j^0 - u^0(x_j) = 0, \quad j = 1, \dots, n. \quad (2.22)$$

In the next step we rewrite (2.21)-(2.22) in the form of (1.4). To this end we define the vector space of the grid functions $\mathcal{K}_M$, defined on the grid points $x_m : 1 \leq m \leq M$. If we consider u_m^k for the time level t_k for each k, then the denoted vector is $\mathbf{u}^k \in \mathcal{K}_M$. The operators φ_n, ψ_n in Definition 1.1.3 are defined as the grid restriction operators. Hence, (2.21)-(2.22) can be written as

$$\frac{\mathbf{u}^{k+1} - \mathbf{u}^k}{\tau} - (1-\theta)D_p^2 \mathbf{u}^k - \theta D_p^2 \mathbf{u}^{k+1} = 0, \quad k = 0, \dots, K-1, \quad (2.23)$$

$$\mathbf{u}^0 - \varphi_n(u^0) = 0, \quad (2.24)$$

where $\mathbf{u}^0 = (u^0(x_1), \dots, u^0(x_M)) \in \mathcal{K}_M$ and $D_p^2 \in \mathbb{R}^{M\times M}$ denotes the standard discretization matrix of the second derivative with periodic boundary conditions, i.e.,

$$D_p^2 = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix}.$$

We choose the discrete normed spaces as $X_n = Y_n = \underbrace{\mathcal{K}_M \times \ldots \times \mathcal{K}_M}_{K+1}$, hence $\mathbf{v}_n := (\mathbf{v}^0, \dots, \mathbf{v}^K) \in X_n$. We introduce the following norms:

- ◇ in $\mathcal{K}_M$: $\|\mathbf{v}^k\|_{\mathcal{K}_M} = \max\limits_{1\le m\le M} |v^k(x_m)| = \|\mathbf{v}^k\|_\infty$,
- ◇ in X_n: $\|\mathbf{v}_n\|_{X_n} = \max\limits_{0\le k\le K} \|\mathbf{v}^k\|_{\mathcal{K}_M}$,
- ◇ in Y_n: $\|\mathbf{v}_n\|_{Y_n} = \|\mathbf{v}^0\|_{\mathcal{K}_M} + \sum\limits_{k=1}^{K} \tau \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$.

Let $\mathbf{v}_n \in X_n$ be any element and we denote by $\eta_n = (\eta^0, \dots, \eta^K) \in Y_n$ its image. Then the mapping $F_n : X_n \to Y_n$ can be written as $F_n(\mathbf{v}_n) = \eta_n$. Particularly, for our discretization (2.23)-(2.24) it yields the relation

$$\frac{\mathbf{v}^{k+1} - \mathbf{v}^k}{\tau} - (1-\theta)D_p^2\mathbf{v}^k - \theta D_p^2\mathbf{v}^{k+1} = \eta^{k+1},\ k = 0, \dots, K-1,$$

$$\mathbf{v}^0 = \eta^0.$$

Hence, the investigated method can be rewritten in the form

$$Q_1\mathbf{v}^{k+1} = Q_2\mathbf{v}^k + \tau\eta^{k+1}, \tag{2.25}$$

where $Q_1 = I - \theta\tau D_p^2$ and $Q_2 = I + (1-\theta)\tau D_p^2$ are the subtransition matrices (which depend on h and τ). Introducing the notation $r = \tau/h^2$ we can write Q_1 as

$$Q_1 = \begin{pmatrix} 1{+}2r\theta & -r\theta & 0 & \cdots & 0 & 0 & -r\theta \\ -r\theta & 1{+}2r\theta & -r\theta & 0 & \cdots & 0 & 0 \\ 0 & -r\theta & 1{+}2r\theta & -r\theta & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -r\theta & 1{+}2r\theta & -r\theta & 0 \\ 0 & \cdots & \cdots & 0 & -r\theta & 1{+}2r\theta & -r\theta \\ -r\theta & 0 & 0 & \cdots & 0 & -r\theta & 1{+}2r\theta \end{pmatrix}.$$

Since $r > 0$ and $\theta \in [0,1]$, Q_1 is strictly diagonally dominant and $(Q_1)_{ij} \le 0$ for all $i \ne j$. Hence, Q_1 is an M-matrix with the dominating vector $g = (1, \dots, 1)^T$. Due to a basic result corresponding to M-matrices (see e.g. [11]), we have the estimate

$$\left\|Q_1^{-1}\right\|_\infty \le \frac{\|g\|_\infty}{\min\limits_{1\le i\le M}(Q_1 g)_i} = \frac{1}{1{+}2r\theta{-}r\theta{-}r\theta} = 1. \tag{2.26}$$

The matrix Q_2 can be written in the form

$$Q_2 = \begin{pmatrix} a & b & 0 & \cdots & 0 & 0 & b \\ b & a & b & 0 & \cdots & 0 & 0 \\ 0 & b & a & b & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & b & a & b & 0 \\ 0 & \cdots & \cdots & 0 & b & a & b \\ b & 0 & 0 & \cdots & 0 & b & a \end{pmatrix},$$

where $a = 1-2r(1-\theta)$ and $b = r(1-\theta)$. It is easy to see that under the assumption $r \leq 1/[2(1-\theta)]$ we have

$$\|Q_2\|_\infty = 1. \tag{2.27}$$

Introducing the notation $Q = Q_1^{-1}Q_2$, the iteration (2.25) can be written as

$$\mathbf{v}^{k+1} = Q\mathbf{v}^k + \tau Q_1^{-1}\eta^{k+1}.$$

Applying the above recursion and putting $\mathbf{v}^0 = \eta^0$ for any $k = 0, 1, \ldots, K$, we get

$$\mathbf{v}^k = Q^k\mathbf{v}^0 + \sum_{j=1}^{k} \tau Q^{j-1}Q_1^{-1}\eta^{k+1-j} = Q^k\eta^0 + \tau Q^{k-1}Q_1^{-1}\eta^1 + \ldots + \tau Q_1^{-1}\eta^k.$$

Hence, according to the introduced norms, we obtain the estimate

$$\|\mathbf{v}_n\|_{X_n} \leq \max_{0\leq k\leq K} \left\{ ||Q^k||_\infty \max_{1\leq j\leq k} \left\{ ||Q^{j-1}Q_1^{-1}||_\infty \right\} \right\} \|\eta_n\|_{Y_n}. \tag{2.28}$$

Using the relations (2.26) and (2.27), we get that

$$||Q||_\infty = ||Q_1^{-1}Q_2||_\infty \leq ||Q_1^{-1}||_\infty ||Q_2||_\infty \leq 1,$$

thus $||Q^k||_\infty \leq ||Q||_\infty^k \leq 1$ and similarly $||Q^{j-1}Q_1^{-1}||_\infty \leq ||Q||_\infty^{j-1}||Q_1^{-1}||_\infty \leq 1$. Hence, obviously we have

$$\max_{0\leq k\leq K} \left\{ ||Q^k||_\infty \max_{1\leq j\leq k} \left\{ ||Q^{j-1}Q_1^{-1}||_\infty \right\} \right\} = 1.$$

Since $F_n(\mathbf{v}_n) = \eta_n$, we can rewrite (2.28) as

$$\|\mathbf{v}_n\|_{X_n} \leq \|\eta_n\|_{Y_n} = \|F_n(\mathbf{v}_n)\|_{Y_n}. \tag{2.29}$$

For any elements $\mathbf{z}_n, \mathbf{w}_n \in X_n$ we denote by ϱ_n and ξ_n their image, i.e., $F_n(\mathbf{z}_n) = \varrho_n$ and $F_n(\mathbf{w}_n) = \xi_n$. This results in the relations

$$\frac{\mathbf{z}^{k+1} - \mathbf{z}^k}{\tau} - (1-\theta)D_p^2\mathbf{z}^k - \theta D_p^2\mathbf{z}^{k+1} = \varrho^{k+1}, \ k = 0, \ldots, K-1, \tag{2.30}$$

$$\mathbf{z}^0 - \varphi_n(u^0) = \varrho^0,$$

$$\frac{\mathbf{w}^{k+1} - \mathbf{w}^k}{\tau} - (1-\theta)D_p^2\mathbf{w}^k - \theta D_p^2\mathbf{w}^{k+1} = \xi^{k+1}, \ k = 0, \ldots, K-1, \tag{2.31}$$

$$\mathbf{w}^0 - \varphi_n(u^0) = \xi^0.$$

Subtracting (2.30) from (2.31), we gain

$$\mathbf{z}^{k+1} - \mathbf{w}^{k+1} = Q(\mathbf{z}^k - \mathbf{w}^k) + \tau Q_1^{-1}(\varrho^{k+1} - \xi^{k+1}), \quad k = 0, \ldots, K-1.$$

Using (2.29) by the notation $\mathbf{v}_n = \mathbf{z}_n - \mathbf{w}_n$, we obtain

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} \leq \|\varrho_n - \xi_n\|_{Y_n} = \|F_n(\mathbf{z}_n) - F_n(\mathbf{w}_n)\|_{Y_n}.$$

It is easy to see that the above estimation is in the form of (2.2) with $C = 1$.

Theorem 2.3.1. *Under the condition $r \leq 1/[2(1-\theta)]$ the θ-method is N-stable in the introduced norm for the periodic initial-value diffusion problem (2.18)-(2.20).*

Remark 2.3.3. In the maximum norm θ-method's consistency order is one both in time and space (in case of $\theta \neq 0$, see for further details [60] Example 2.4.3). If $\theta = 1/2$, then the order of consistency in space is two.

Theorem 2.3.2. *Under the condition $r \leq 1/[2(1-\theta)]$ the θ-method is convergent in the introduced norm for the periodic initial-value diffusion problem (2.18)-(2.20).*

Proof. Using Remark 2.3.3 and Theorem 2.3.1, from Theorem 2.0.1 we immediately get the result. ■

Reaction-diffusion problem

Further we consider the following periodic initial-value reaction-diffusion problem:

$$\partial_t u(t,x) = \partial_{xx} u(t,x) + f(u), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.32}$$
$$u(t,x) = u(t,x+1), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.33}$$
$$u(0,x) = u^0(x), \qquad x \in \mathbb{R}, \tag{2.34}$$

where $T \in \mathbb{R}^+$ and in equation (2.32) we assume that $f : \mathbb{R} \to \mathbb{R}$ is a given globally Lipschitz continuous reaction function. The conditions (2.33)-(2.34) are periodic boundary conditions and initial-value conditions, where u^0 is a given one-periodic function. It is easy to see that the continuous problem (2.32)-(2.34) can be rewritten in the form of (1.1). As we have mentioned, we assume the existence of the unique, sufficiently smooth solution of the problem (2.32)-(2.34).

Remark 2.3.4. Since the solution is periodic, it is sufficient to determine the solution in one period only.

Let us take the formerly introduced spatial and time grids and norms. We apply the introduced θ-method based on the previous train of thought.
For any elements $\mathbf{z}_n, \mathbf{w}_n \in X_n$ we denote by ϱ_n and ξ_n their image, i.e., $F_n(\mathbf{z}_n) = \varrho_n$ and $F_n(\mathbf{w}_n) = \xi_n$. Then we consider the following two problems:

$$\frac{\mathbf{z}^{k+1} - \mathbf{z}^k}{\tau} - (1-\theta) D_p^2 \mathbf{z}^k - \theta D_p^2 \mathbf{z}^{k+1} - \mathbf{f}(\mathbf{z}^k) = \varrho^{k+1}, k = 0, \ldots, K-1, \tag{2.35}$$

$$\mathbf{z}^0 - \varphi_n(u^0) = \varrho^0,$$

$$\frac{\mathbf{w}^{k+1}-\mathbf{w}^k}{\tau}-(1-\theta)D_p^2\mathbf{w}^k-\theta D_p^2\mathbf{w}^{k+1}-\mathbf{f}(\mathbf{w}^k)=\xi^{k+1}, k=0,\dots,K-1, \quad (2.36)$$
$$\mathbf{w}^0-\varphi_n(u^0)=\xi^0,$$

where $\mathbf{f}$ denotes the grid function defined on the grid points x_m, so it means that $[\mathbf{f}(\mathbf{z}^k)]_m=\varphi_n(f(x_m))$ for all $m=1,\dots,M$. Subtracting (2.36) from (2.35), for $k=0,\dots,K-1$, we get

$$\mathbf{z}^{k+1}-\mathbf{w}^{k+1}=Q_1^{-1}Q_2(\mathbf{z}^k-\mathbf{w}^k)+\tau Q_1^{-1}(\mathbf{f}(\mathbf{z}^k)-\mathbf{f}(\mathbf{w}^k))+\tau Q_1^{-1}(\varrho^{k+1}-\xi^{k+1}), \quad (2.37)$$

where Q_1 and Q_2 are the earlier defined subtransition matrices. Since f is a given globally Lipschitz continuous function, this implies

$$\|\mathbf{f}(\mathbf{z}^k)-\mathbf{f}(\mathbf{w}^k)\|_{\mathcal{K}_M}\le L\|\mathbf{z}^k-\mathbf{w}^k\|_{\mathcal{K}_M}. \quad (2.38)$$

Using the Lipschitz property (2.38) and applying the results (2.26) and (2.27) respectively, we get $\left\|Q_1^{-1}Q_2\right\|_\infty\le 1$. Then recursion (2.37) shows us that for $k=0,\dots,K$ we have the estimate

$$\|\mathbf{z}^k-\mathbf{w}^k\|_{\mathcal{K}_M}\le(1+\tau L)\|\mathbf{z}^{k-1}-\mathbf{w}^{k-1}\|_{\mathcal{K}_M}+\tau\|\varrho^k-\xi^k\|_{\mathcal{K}_M}. \quad (2.39)$$

Applying recursion (2.39) and $\mathbf{z}^0-\mathbf{w}^0=\varrho^0-\xi^0$, we get for $k=0,\dots,K$

$$\|\mathbf{z}^k-\mathbf{w}^k\|_{\mathcal{K}_M}\le(1+\tau L)^k\|\varrho_n-\xi_n\|_{Y_n}.$$

Using that $\tau K=T$, we get the estimation

$$\|\mathbf{z}_n-\mathbf{w}_n\|_{X_n}\le(1+\tau L)^K\|\varrho_n-\xi_n\|_{Y_n}\le e^{LT}\|F_n(\mathbf{z}_n)-F_n(\mathbf{w}_n)\|_{Y_n}.$$

It is easy to see that the above estimation is in the form of (2.2) with $C=e^{LT}$.

Theorem 2.3.3. *Under the condition $r\le 1/[2(1-\theta)]$, for the Lipschitzian forcing term f the θ-method is N-stable in the introduced norm for the periodic initial-value reaction-diffusion problem* (2.32)-(2.34).

Remark 2.3.5. Due to these results we automatically get that

(a) both the explicit finite difference method (for $\theta=0$) under the condition $r\le 1/2$ and the implicit method finite difference method (for $\theta=1$) without any condition are N-stable,

(b) if the given forcing term is in the form of $f(t,u)$ and it is a Lipschitz continuous function with respect to its second variable by the constant L, then we can similarly verify that the θ-method is convergent for the investigated problem.

As we can see using the N-stability notion we obtained the well-known stability results. It has been summarized in Table 2.5.

method		complexity	stability	convergence	
θ	name	explicit/implicit	$r = \tau/h^2$	time	space
0	forward Euler	explicit	$r \leq 0.5$	1	1
1	backward Euler	implicit	–	1	1
0.5	Crank–Nicolson	implicit	$r \leq 1$	1	2
θ	θ-method	explicit/implicit	$r \leq 1/2(1-\theta)$	1	1 or 2

Table 2.5. *The N-stability properties to diffusion problems.*

2.3.2 Transport problems

In the sequel, we apply the N-stability technique to verify the stability of hyperbolic equations, namely, to the periodic initial-value transport problem. We consider the problem

$$\partial_t u(t,x) + a\partial_x u(t,x) = 0, \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.40}$$

$$u(t,x) = u(t,x+1), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.41}$$

$$u(0,x) = u^0(x), \qquad x \in \mathbb{R}, \tag{2.42}$$

where $T \in \mathbb{R}^+$ and $a \in \mathbb{R}$ are fixed constants. The conditions (2.41)-(2.42) are periodic boundary conditions and initial-value conditions, where u^0 is a given one-periodic function. The periodic boundary condition appears in the stability investigation of the "good" Boussinesq equation in [49].

One can see that the continuous problem (2.40)-(2.42) can be rewritten in the form (1.1). Let $u^0(x) \in C^1(\mathbb{R})$ be a given function, then the problem (2.40)-(2.42) has the unique solution in the form $u(x,t) = u^0(x-at)$. Since the solution is periodic, it is sufficient to determine it on one period only. We define both the spatial and time grids, as follows. The spatial grid points are

$$\{x_m = mh,\ \text{where } m = 1,\dots,M,\ h = 1/M \text{ and } M \in \mathbb{N},\ M \geq 2\},$$

and the time levels are

$$\{t_k = k\tau,\ \text{where } k = 0,\dots,K \text{ and } \tau = T/K\}.$$

Applying the centralized Crank–Nicolson-method to this transport problem, for $m = 1,\dots,M$ and $k = 0,\dots,K-1$ we gain the numerical scheme as follows

$$u_m^{k+1} + \frac{\tau a}{4h}\left(u_{m+1}^{k+1} - u_{m-1}^{k+1}\right) = u_m^k - \frac{\tau a}{4h}\left(u_{m+1}^k - u_{m-1}^k\right). \tag{2.43}$$

Using the periodic boundary conditions, we put $u_0^{k-1} = u_M^{k-1}$, $u_1^{k-1} = u_{M+1}^{k-1}$ and $u_0^{k+1} = u_M^{k+1}$, $u_1^{k+1} = u_{M+1}^{k+1}$. The discretization of the initial-value condition can be written as

$$u_m^0 - u^0(x_m) = 0, \quad m = 1,\dots,M. \tag{2.44}$$

In the next step we rewrite (2.43)-(2.44) in the form (1.4). Similarly to the reaction-diffusion problem, we define the vector space of the grid functions $\mathcal{K}_M$, defined at the grid points $x_m : 1 \leq m \leq M$. If we consider u_m^k for the time level t_k for each k, then the denoted vector is $\mathbf{u}^k \in \mathcal{K}_M$. We define the mappings φ_n and ψ_n as grid functions.

Introducing the notation $R = a\tau/h$ the equations (2.43)-(2.44) can be written as

$$\mathbf{u}^{k+1} + D_p\mathbf{u}^{k+1} = \mathbf{u}^k - D_p\mathbf{u}^k, \ k = 0, \ldots, K-1, \tag{2.45}$$

$$\mathbf{u}^0 - \varphi_n(u^0) = 0, \tag{2.46}$$

where $\mathbf{u}^0 = (u^0(x_1), \ldots, u^0(x_M)) \in \mathcal{K}_M$ and $D_p \in \mathbb{R}^{M\times M}$ denotes the standard discretization matrix with periodic boundary conditions, i.e.,

$$D_p = \begin{pmatrix} 0 & \frac{R}{4} & 0 & \cdots & 0 & 0 & -\frac{R}{4} \\ -\frac{R}{4} & 0 & \frac{R}{4} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{R}{4} & 0 & \frac{R}{4} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\frac{R}{4} & 0 & \frac{R}{4} & 0 \\ 0 & \cdots & \cdots & 0 & -\frac{R}{4} & 0 & \frac{R}{4} \\ \frac{R}{4} & 0 & 0 & \cdots & 0 & -\frac{R}{4} & 0 \end{pmatrix}. \tag{2.47}$$

Using the notations $Q_1 = (I + D_p)$ and $Q_2 = (I - D_p)$, respectively, the discretization (2.45)-(2.46) yields the problem

$$Q_1\mathbf{u}^{k+1} = Q_2\mathbf{u}^k, \ k = 0, \ldots, K-1, \tag{2.48}$$

$$\mathbf{u}^0 = \varphi_n(u^0). \tag{2.49}$$

To prove the existence of the inverse of Q_1, we use the fact that D_p is a skew-symmetric matrix. Therefore its eigenvalues are on the imaginary axes, hence $Q_1 = (I + D_p)$ has no zero eigenvalue and therefore it is regular. Then, we can rewrite (2.48)-(2.49) as

$$\mathbf{u}^{k+1} = Q_1^{-1}Q_2\mathbf{u}^k, \ k = 0, \ldots, K-1,$$
$$\mathbf{u}^0 = \varphi_n(u^0).$$

Let the normed spaces be $X_n = Y_n = \underbrace{\mathcal{K}_M \times \ldots \times \mathcal{K}_M}_{K+1}$, so $\mathbf{v}_n := (\mathbf{v}^0, \ldots, \mathbf{v}^K) \in X_n$.

We define the mapping $F_n : X_n \to Y_n$ on any element $\mathbf{v}_n \in X_n$ as

$$[F_n(\mathbf{v}_n)]_k = \begin{cases} \mathbf{v}^0 - \varphi_n(u^0), & k = 0, \\ \mathbf{v}^k - Q_1^{-1}Q_2\mathbf{v}^{k-1}, & k = 1, 2, \ldots, K \end{cases} \tag{2.50}$$

where $\mathbf{v}^0 := \mathbf{v}^0 - \varphi_n(u^0)$. From the relation (2.50) we can express $\mathbf{v}^k$ as

$$\mathbf{v}^k = \begin{cases} [F_n(\mathbf{v}_n)]_0 + \varphi_n(u^0), & k = 0, \\ [F_n(\mathbf{v}_n)]_k + Q_1^{-1}Q_2\mathbf{v}^{k-1}, & k = 1, 2, \ldots, K. \end{cases} \tag{2.51}$$

The first step is to prove the N-stability property. To this aim we define the norm in $\mathcal{K}_M$ as

$$\|\mathbf{v}^k\|_{\mathcal{K}_M} = h\left(\sum_{m=1}^{M} |v^k(x_m)|^2\right)^{1/2} = \left\|\mathbf{v}^k\right\|_2 .$$

Before we start to prove the N-stability property, we give a useful norm relation which helps us how to choose properly the norms in X_n and Y_n, respectively. Using (2.51) and the exact value of $\left\|Q_1^{-1}Q_2\right\|_2$ (see Lemma A.2.3) for $i = 0, \dots, K$ we get

$$\begin{aligned}
\sum_{k=0}^{i} \left\|\mathbf{z}^k - \mathbf{w}^k\right\|_2 &\le \sum_{k=0}^{i} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 + \sum_{k=0}^{i-1} \left\|\mathbf{z}^k - \mathbf{w}^k\right\|_2 \\
&\le \sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 + \sum_{k=0}^{i-1} \left\|\mathbf{z}^k - \mathbf{w}^k\right\|_2 \\
&\le 2\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 + \sum_{k=0}^{i-2} \left\|\mathbf{z}^k - \mathbf{w}^k\right\|_2 \\
&\le \dots \\
&\le K\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 + \left\|\mathbf{z}^0 - \mathbf{w}^0\right\|_2 \\
&= K\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 + \|[F_n(\mathbf{z}_n)]_0 - [F_n(\mathbf{w}_0)]_k\|_2 \\
&\le (K+1)\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 \\
&\le 2K\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 .
\end{aligned}$$

Hence, for $i = 0, \dots, K$ we obtain the estimate

$$\sum_{k=0}^{i} \left\|\mathbf{z}^k - \mathbf{w}^k\right\|_2 \le 2K\sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_2 . \tag{2.52}$$

We introduce the following norms in X_n and Y_n:

$$\begin{aligned}
\text{in } X_n : \|\mathbf{v}_n\|_{X_n} &= \tau\sum_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M} , \\
\text{in } Y_n : \|\mathbf{v}_n\|_{Y_n} &= \sum_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M} .
\end{aligned} \tag{2.53}$$

Then, based on (2.52) we get

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} \le 2T \|F_n(\mathbf{z}_n) - F_n(\mathbf{w}_n)\|_{Y_n} ,$$

where $\tau K = T$.

One can see that the above estimate is in the form of (2.2) with $C = 2T$. Therefore we proved the validity of the following statement.

Theorem 2.3.4. *The centralized Crank–Nicolson-method is N-stable for the periodic initial-value transport problem* (2.40)-(2.42) *in the norm* (2.53).

Remark 2.3.6. In the norm (2.53) the order of consistency of the centralized Crank–Nicolson-method is two both in time and space (see for further details [60] Section 5.4.4).

Theorem 2.3.5. *The centralized Crank–Nicolson-method is convergent for the periodic initial-value transport problem* (2.40)-(2.42) *and the order of convergence is two both in time and space.*

Proof. Using Remark 2.3.6 and Theorem 2.3.4, then from Theorem 2.0.1 we immediately get the result. ■

Remark 2.3.7. Based on estimate (2.52) we are able to define N-stability for other norms, too.

- ◇ The $C = 2T$ stability constant of (2.2) can be reached if we define the norms as
  - in X_n: $\|\mathbf{v}_n\|_{X_n} = \max\limits_{0\le k\le K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$,
  - in Y_n: $\|\mathbf{v}_n\|_{Y_n} = \sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$.
- ◇ Using the relation
$$\sqrt{\sum_{k=0}^{K} \|\mathbf{v}^k\|_{\mathcal{K}_M}^2} \le \sum_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$$
we can define the following norms:
  - in X_n: $\|\mathbf{v}_n\|_{X_n} = \tau \left(\sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}^2 \right)^{1/2}$,
  - in Y_n: $\|\mathbf{v}_n\|_{Y_n} = \sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$.

In this case the N-stability estimation (2.2) is valid with $C = 2T$.

Remark 2.3.8. Consider the relation (2.51). For the first term we can give the following estimate for $i = 0, \dots, K$

$$\left\|\mathbf{z}^i - \mathbf{w}^i\right\|_{\mathcal{K}_M} \le \|[F_n(\mathbf{z}_n)]_i - [F_n(\mathbf{w}_n)]_i\|_{\mathcal{K}_M} + \left\|\mathbf{z}^{i-1} - \mathbf{w}^{i-1}\right\|_{\mathcal{K}_M}.$$

Applying this iteration for $i = 0, \dots, K$, we gain the estimation

$$\left\|\mathbf{z}^i - \mathbf{w}^i\right\|_{\mathcal{K}_M} \le \sum_{k=0}^{i} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_{\mathcal{K}_M} \le \sum_{k=0}^{K} \|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\|_{\mathcal{K}_M}.$$

Hence, we can define two more norms for which the N-stability property holds.

- By choosing the norms as
  - in X_n: $\|\mathbf{v}_n\|_{X_n} = \max\limits_{0\le k\le K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$,
  - in Y_n: $\|\mathbf{v}_n\|_{Y_n} = \sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$,

  and due to the obvious relation

$$\max_{0\le k\le K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M} \le \sum_{k=0}^{K} \|[F_n(\mathbf{v}_n)]_k\|_{\mathcal{K}_M},$$

  we get the stability constant $C = 1$.

- By defining the norms as
  - in X_n: $\|\mathbf{v}_n\|_{X_n} = \tau \max\limits_{0\le j\le K} \left\|\mathbf{v}^j\right\|_{\mathcal{K}_n}$,
  - in Y_n: $\|\mathbf{v}_n\|_{Y_n} = \max\limits_{0\le k\le K} \|[F_n(\mathbf{v}_n)]_k\|_{\mathcal{K}_M}$

  and since the relation

$$\max_{0\le k\le K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M} \le K \max_{0\le k\le K} \|[F_n(\mathbf{v}_n)]_k\|_{\mathcal{K}_M}$$

  holds, we have N-stability with $C = 1$.

Transport problem with forcing term

Consider the periodic initial-value transport problem with forcing term, i.e.,

$$\partial_t u(t,x) + a\partial_x u(t,x) = f(t,x), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.54}$$
$$u(t,x) = u(t,x+1), \qquad x \in \mathbb{R},\ t \in [0,T], \tag{2.55}$$
$$u(0,x) = u^0(x), \qquad x \in \mathbb{R}, \tag{2.56}$$

where $T \in \mathbb{R}^+$ and $a \in \mathbb{R}$ are fixed constants and $f(t,x)$ is a given function. For this problem we get the equalities

$$\mathbf{u}^{k+1} - Q_1^{-1}Q_2\mathbf{u}^k = Q_1^{-1}\mathbf{f}^{k+1},\ k = 0,\dots,K-1,$$
$$\mathbf{u}^0 = \varphi_n(u^0),$$

where $\mathbf{f}$ denotes the grid function defined on the grid points x_m for all $m = 1,\dots,M$. Considering the mapping (2.50) and introducing the element $\mathbf{g}_n \in Y_n$ to the right-hand side we can define the F_n^g operator to the problem (2.54)-(2.56) as

$$F_n^g(\mathbf{v}_n) = F_n(\mathbf{v}_n) - \mathbf{g}_n.$$

We can prove the N-stability property as follows. Let us suppose that Theorem 2.3.4, i.e. the centralized Crank–Nicolson-method is N-stable for the problem (2.40)-(2.42), then for arbitrary $\mathbf{z}_n, \mathbf{w}_n \in X_n$ the relation

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} \le C\,\|F_n^g(\mathbf{z}_n) - F_n^g(\mathbf{w}_n)\|_{Y_n}$$

holds. Since $F_n^g(\mathbf{v}_n) = F_n(\mathbf{v}_n) - \mathbf{g}_n$, we can rewrite the above estimation as

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} \leq C \|F_n(\mathbf{z}_n) - \mathbf{g}_n - F_n(\mathbf{w}_n) + \mathbf{g}_n\|_{Y_n} = C \|F_n(\mathbf{z}_n) - F_n(\mathbf{w}_n)\|_{Y_n} .$$

Then, according to Section 2.3.2, the stability relation holds with $C = 2T$. Thus, the periodic initial-value transport problem with forcing term is N-stable, too.

Theorem 2.3.6. *The centralized Crank–Nicolson-method is N-stable for the periodic initial-value transport problem with forcing term of the form (2.54)-(2.56) in the norm (2.53).*

As we could see, the N-stability notion is useful from the application point of view. To prove this property the key point is the proper definition of the φ_n and ψ_n mappings, the normed spaces of the discrete problems and the corresponding norms. It has been summarized in Table 2.6.

	Reaction-diffusion problem	Transport problem
$\varphi_n,\ \psi_n$	grid functions	grid functions
$\mathcal{K}_n$	VS of grid functions	VS of grid functions
$X_n \equiv Y_n$	$\underbrace{\mathcal{K}_M \times \ldots \times \mathcal{K}_M}_{K+1}$	$\underbrace{\mathcal{K}_M \times \ldots \times \mathcal{K}_M}_{K+1}$
$\left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$	$\max\limits_{1\leq j\leq M} \lvert v^k(x_j)\rvert$	$h\left(\sum\limits_{j=1}^{M} \lvert v^k(x_j)\rvert^2\right)^{1/2}$
$\|\mathbf{v}_n\|_{X_n}$	$\max\limits_{0\leq k\leq K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$	$\tau \sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$
$\|\mathbf{v}_n\|_{Y_n}$	$\|\mathbf{v}^0\|_{\mathcal{K}_M} + \sum\limits_{k=1}^{K} \tau \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$	$\sum\limits_{k=0}^{K} \left\|\mathbf{v}^k\right\|_{\mathcal{K}_M}$

Table 2.6. *How to choose operators, normed spaces and corresponding norms to prove N-stability.*

2.4 Evolution equations

This section deals with the connection between the introduced framework with N-stability and evolution equations.

2.4.1 Equivalence theorem for linear evolution equations

In the paper [54] the authors extended the classical Lax–Richtmyer equivalence theorem for linear operator equations, so the classical result under certain assumptions is true not only for initial-value problems, but also for boundary value and mixed problems. The theory relies on Stetter's framework and on the generalized Banach–Steinhaus theorem which is proved in [51].

Here we briefly summarize their results and show the connection between their theorem and Lax and Richtmyer's theory for linear evolution equation.

Setting the problem

Let us consider the problem

$$Lu = f, \quad f \in Y, \tag{2.57}$$

where $(X, ||\cdot||_X)$ and $(Y, ||\cdot||_Y)$ are normed spaces, $\mathrm{dom}(L) \subset X$ and $\mathrm{ran}(L) \subset Y$. One can see that (2.57) is in the form of (1.1). Furthermore, we suppose the (2.57) is well-posed in the following sense: $\overline{\mathrm{ran}(L)} = Y$, i.e. the range of the linear operator L is dense in Y and there exists a bounded linear operator $E \in B(Y, X)$ such that the composition EL is the identity in $\mathrm{dom}(A)$.

Remark 2.4.1. This well-posedness condition implies that the equation (2.57) for $f \in \mathrm{ran}(Y)$ has the unique solution $u^* = Ef \in \mathrm{dom}(L)$. When $f \in Y \setminus \mathrm{ran}(Y)$, then there is no solution. Based on the results of paper [50] operator E is the unique bounded extension to Y of operator $L^{-1} : \mathrm{ran}(L) \to \mathrm{dom}(L)$, therefore in this case Ef can be regarded as a generalized solution.

Let us consider the sequence of the discrete problems

$$L_n u_n = f_n, \quad f_n \in Y_n, \ n \in \mathbb{I}, \tag{2.58}$$

where $(X_n, ||\cdot||_{X_n})$ and $(Y_n, ||\cdot||_{Y_n})$ are normed spaces, $L_n : X_n \to Y_n$ is a linear operator for all $n \in \mathbb{I}$. One can see that (2.58) is in the form of (1.4). We assume the well-posedness of problem (2.58) similarly to the previous case with the solution operator $E_n = L_n^{-1}$.

In order to make a connection between the problems (2.57) and (2.58) we give assumptions for the mappings φ_n and ψ_n.

Assumption 2.4.1. For the mappings φ_n and ψ_n from Definition 1.1.3 we require the following:

(a) $\varphi_n \in B(X, X_n)$ and $\psi_n \in B(Y, Y_n)$ for all $n \in \mathbb{I}$,

(b) $||\varphi_n||_{B(X,X_n)} \leq C_1$ and $||\psi_n||_{B(Y,Y_n)} \leq C_2$ for all $n \in \mathbb{I}$, where the constants C_1 and C_2 are independent of n,

(c) $\psi_n(f) = f_n$.

Remark 2.4.2. The conditions of Assumption 2.4.1 is not too restrictive. For instance in case of projections these are fulfilled.

Taking into account the introduced setting we can imagine the general discretization process as in Figure 2.1.

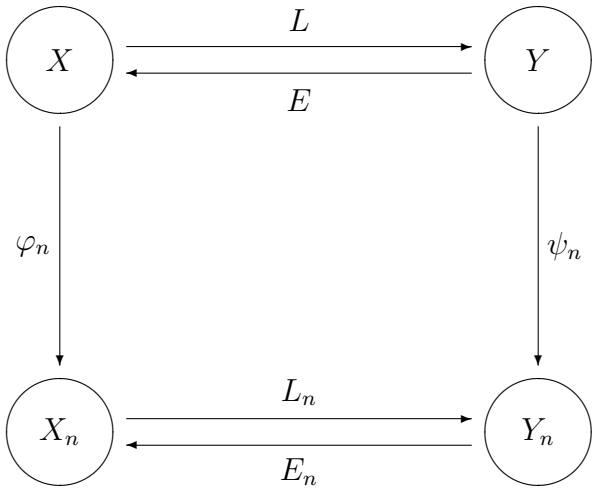


Figure 2.1. *The general discretization process based on the generalized equivalence theorem.*

Basic notions

Based on Section 2.3 of paper [54] we give the basic notions for problem (2.57).

Definition 2.4.1. *The family* $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ *is called a method for the solution of problem* (2.57).

Definition 2.4.2.

- ◇ *Let* F *be an element in* Y. *We say that the method* $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ *is convergent for the problem* (2.57) *if the relation*

$$\lim_{n\to\infty} ||\varphi_n(Ef) - E_n\psi_n(f)||_{X_n} = 0 \tag{2.59}$$

  holds. We say that the method is convergent if it is convergent for each problem (2.57) *as* f *ranges in* Y.

- ◇ *Let* v *be a given element in* $\mathrm{dom}(L)$. *We say that the method is consistent on the element* v *if the relation*

$$\lim_{n\to\infty} ||L_n\varphi_n(v) - \psi_n(Lv)||_{Y_n} = 0 \tag{2.60}$$

  holds. A method $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ *is consistent if it is consistent at each* v *in a set* L^* *such that the image* $L(L^*)$ *is dense in* Y.

- ◇ *The method* $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ *is stable if there exists a positive constant* C *independent of* n *such that*

$$||E_n||_{B(Y_n,X_n)} \leq C. \tag{2.61}$$

General equivalence theorem

Before proving the general Lax equivalence theorem we mentioned an important lemma which is called the generalized Banach–Steinhaus theorem by Palencia and Sanz-Serna.

Lemma 2.4.1 (Section 2, Lemma, [51]). *Let $\mathscr{Y}$ be a Banach space, $(Y_n)_{n\in\mathbb{I}}$ a family of normed spaces, $T_n : \mathscr{Y} \to Y_n$ linear operators. If for each $v \in \mathscr{Y}$, $\sup ||T_n v||_{Y_n} < \infty$, then $\sup ||T_n||_{B(\mathscr{Y},Y_n)} < \infty$.*

Theorem 2.4.2 (Section 2.4, Thm. 1, [54]). *Let us consider the operator L and the normed spaces X and Y and the method $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$. Furthermore, we suppose that Assumption 2.4.1 is fulfilled.*

i, If the method $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ is consistent and stable, then it is convergent in sense of Definition 2.4.2.

ii, If the method $(X_n, Y_n, L_n, \varphi_n, \psi_n)_{n\in\mathbb{I}}$ is convergent, then it is stable in sense of Definition 2.4.2 provided that Y is a Banach space and the following condition holds:

[C] There exists a constant K such that for each $g \in Y_n$ with $||g|| < 1$, there exists an element $f \in Y$ such that $||f|| < K$ and $\psi_n(f) = g$.

Proof. We will prove the theorem in two steps.

$i, \to ii$, Let $f \in L(L^*)$. The convergence for the problem (2.57) follows from (2.60) and (2.61), since

$$\begin{aligned} ||\varphi_n(Ef) - E_n\psi_n(f)||_{X_n} &= ||E_n(L_n\varphi_n(u) - \psi_n(Lu))||_{X_n} \\ &\le C||L_n\varphi_n(u) - \psi_n(Lu)||_{Y_n}. \end{aligned}$$

If $f \in Y$, $f \notin L(L^*)$, we can choose a sequence (f^k) with $(f^k) \in L(L^*)$ such that $\lim f^k = f$. Then we have

$$\begin{aligned} ||\varphi_n(Ef) - E_n\psi_n(f)||_{X_n} \le\ & ||\varphi_n(Ef) - \varphi_n(Ef^k)||_{X_n} \\ & + ||\varphi_n(Ef^k) - E_n\psi_n(f^k)||_{X_n} \\ & + ||E_n\psi_n(f^k) - E_n\psi_n(f)||_{X_n}. \end{aligned}$$

Since the operators E, E_n, φ_n, ψ_n can be bounded independently of n, the first and third terms of the right-hand side can be made arbitrarily small, uniformly in n, by taking k large, while the second term tends to zero with n.

$ii, \to i$, Let $f \in Y$. Taking into account Assumption 2.4.1 (b) the norms $||\varphi_n(Ef)||$ are bounded as n tends to ∞. Due to (2.59) we conclude that the norms $||E_n\psi_n(f)||$ are also bounded. Using the generalized Banach–Steinhaus Lemma 2.4.1 there exists a constant $\widetilde{C}$ such that $||E_n\psi_n|| \le \widetilde{C}$. If $g \in Y_n$ with $||g|| \le 1$, then using the condition $[C]$ we can write that

$$||E_n g|| = ||E_n\psi_n(f)|| \le \widetilde{C}K,$$

thus $||E_n|| \le \widetilde{C}K$.

■

Applications of Theorem 2.4.2

In this subsection we give two fundamental classes for operator semigroups which are related to the results of Theorem 2.4.2.

Remark 2.4.3. Basic references in the topic of one-parameter semigroups for linear evolution equations are Engel and Nagel [21], [22] and Pazy [52].

Example 2.4.1 (Example 3.1, [54])**.** Let us consider the well-posed abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}u(t) = Au(t), \quad t \in [0,T], \\ u(0) = u_0 \in \mathscr{X}. \end{cases} \tag{2.62}$$

where A is the generator of a strongly continuous semigroup $\mathcal{S}$ in a Banach space $\mathscr{X}$. This problem can be formulated in the form of (1.1) by choosing

- $\diamond$ X to be the space of continuous mappings from $[0,T]$ into the Banach space $\mathscr{X}$ with the supremum norm,
- $\diamond$ $Y = \mathscr{X}$,
- $\diamond$ the linear operator F is defined on the domain

$$\mathrm{dom}(F) = \{u(\cdot) \in X \mid \tfrac{\mathrm{d}}{\mathrm{d}t}u(t) \text{ exists}, \tfrac{\mathrm{d}}{\mathrm{d}t}u(t) = Au(t),\ t \in [0,T]\}$$

and it acts as

$$u(\cdot) \to (Fu)(\cdot) = u(0).$$

Due to the results of [42] the difference scheme reads as the recursion

$$u_{k+1} = Q(\tau)u_k,\ k = 0, \dots, K-1,$$

where τ is the step-size, $Q(\tau)$ is a bounded linear operator in $\mathscr{X}$ and u_k is the approximation of $u(k\tau)$. Similarly we can rewrite the discrete problems in the form of (1.4).

- $\diamond$ X_n is the $K+1$ copies of the Banach space $\mathscr{X}$ endowed by the norm

$$\|v\|_{X_n} = \sup_{k=0,\dots,K} \|v_k\|_{\mathscr{X}} \quad \text{for all} \quad v \in X_n,$$

- $\diamond$ Y_n is the $K+1$ copies of the Banach space $\mathscr{X}$ endowed by the norm

$$\|v\|_{Y_n} = \sum_{k=0}^{K} \|v_k\|_{\mathscr{X}} \quad \text{for all} \quad v \in Y_n,$$

- $\diamond$ the mapping φ_n is the grid restriction, i.e.

$$\varphi_n(v) = (v(0), v(\tau), \dots, v(N\tau)), \quad \text{for all} \quad v \in X,$$

- $\diamond$ in this case the mapping ψ_n acts as

$$\psi_n(v) = (v, 0, \dots, 0), \quad \text{for all} \quad v \in Y,$$

◇ in this case the linear operator F_n can be represented as

$$F_n = \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ -Q & I & 0 & \dots & 0 \\ 0 & -Q & I & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -Q & I \end{pmatrix}.$$

Then one can see that Theorem 2.4.2 applied with the above presented X, Y, F, $X_n, Y_n, F_n, \varphi_n, \psi_n$ choices is exactly the Lax–Richtymer equivalence theorem, since the additional conditions of the theorem (Y is a Banach space and the generalized Banach–Steinhaus theorem) are automatically fulfilled.

In order to understand the classical Lax stability [42] we have to calculate the inverse of operator F_n which is denoted by E_n. It can be represented as

$$E_n = \begin{pmatrix} I & 0 & 0 & 0 & \dots & 0 \\ Q & I & 0 & 0 & \dots & 0 \\ Q^2 & Q & I & 0 & \dots & 0 \\ Q^3 & Q^2 & Q & I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ Q^K & Q^{K-1} & Q^{K-2} & \dots & Q & I \end{pmatrix}.$$

The norm of the above matrix is $\sup\limits_{0\le k\le K}\{||Q^k||\}$. Therefore, the stability means in this case

$$\sup_{0\le k\le K}\{||Q(\tau)^k|| \mid 0 \le k\tau \le T\} \le \infty. \tag{2.63}$$

In sense of Definition 2.4.2 the method is stable, since the operator E_n is uniformly bounded. Due to Section 2.1 we know that N-stability and Lax stability (2.63) coincide.

♣

The other example is related to the inhomogeneous case with source function f in L^2.

Example 2.4.2 (Example 3.2, [54])**.** Let us consider the well-posed inhomogeneous abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}u(t) = Au(t) + f(t), \quad t \in [0,T], \\ u(0) = u_0 \in \mathscr{X}, \end{cases} \tag{2.64}$$

where the A is the generator of a strongly continuous semigroup $\mathcal{S}$ in a Banach space $\mathscr{X}$ and $f \in L^2([0,T],\mathscr{X})$. This problem can be formulated in the form of (1.1) by choosing

◇ $X = C([0,T],\mathscr{X})$,

◇ $Y = \mathcal{X} \times L^2([0,T],\mathscr{X})$,

◇ the linear operator F is defined on the domain

$$\mathrm{dom}(F) = \{u(\cdot) \in X \mid \tfrac{\mathrm{d}}{\mathrm{d}t}u(t) \text{ exists}, \tfrac{\mathrm{d}}{\mathrm{d}t}u(t) - Au(t) \in L^2([0,T], \mathscr{X})\}$$

and it acts as

$$u(\cdot) \to Fu(\cdot) = \big(u(0), \tfrac{\mathrm{d}}{\mathrm{d}t}u(t)\big).$$

Remark 2.4.4. It is known that problem (2.64) has the solution

$$S(t)u_0 + \int_0^t S(t-s)f(s)\mathrm{d}s,$$

where the semigroup S is generated by generator A.

Due to Mountain [47], we consider the numerical method

$$u_{k+1} = Q(\tau)u_k + \tau f_k, k = 0, \ldots, K-1,$$

where

$$f_k = \frac{1}{\tau}\int_{k\tau}^{(k+1)\tau} f(t)\mathrm{d}t$$

and $Q(\tau)$ is the operator defined in Example 2.4.1. As before, our aim is to determine the form (1.4). The discrete spaces with the endowed norms $(X_n, ||\cdot||_{X_n})$ and $(Y_n, ||\cdot||_{Y_n})$, the operator F_n and the grid restriction φ_n remain exactly the same from Example 2.4.1. For this case we define the mapping ψ_n as

$$\psi_n(v) = (v, \tau f_0, \ldots, \tau f_{K-1}), \quad \text{for all } v \in Y.$$

In this case we regain the Lax stability (2.63), since X_n, Y_n and F_n have not been altered. ♣

Notes on numerical methods for linear abstract Cauchy problems

Next we briefly summarize numerical analysis techniques for the problems (2.62), (2.64) and the well-posed nonautonomous abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}u(t) = A(t)u(t), \\ u(s) = x \in \mathscr{X}, \end{cases} \tag{2.65}$$

where $s \le t \in [0,T]$, $A(t)$ is the generator of a strongly continuous semigroup $\mathcal{S}(t)$ in a Banach space $\mathscr{X}$ (see for further details [48]).

For problem (2.62) one can use spatial and temporal discreatizations approximating the generator A and the semigroup $\mathcal{S}$, respectively. The mostly cited result establishing the convergence of the spatial discretization is the Trotter-Kato theorem [39]. A deep result about the convergence of the rational type temporal discretization is related to Brenner and Thomée [10].

If it is worth decomposing the operator A of the problems (2.62), (2.64) and (2.65) into sum of simpler operators, then operator splitting methods come into the picture. It has a broad and solid literature and it can be applied mainly to different processes of physics such as Schrödinger equations, Hamilton-Jacobi

equations, Navier-Stokes equations and delay equations. One of the earliest results can be related to Bagrinovskii and Godunov [6], Trotter [63] and Strang [57]. The most widely used splitting methods are Strang [57], Trotter-Lie [63], weighted Trotter-Lie [20] and iterated splitting methods [26]. One can couple splitting methods with spatial-temporal discretizations [8], [7].

The above mentioned techniques are general ones, so these can be appropriately applied to the problems (2.64) and (2.65), too. However, one would like to use specific methods. Exponential integrators [36] and Magnus expansion [46] are for a similar problem to (2.64) and problem (2.65), respectively.

2.4.2 Nonlinear evolution equations

This section deals with discretization methods for nonlinear operator equations written as abstract nonlinear evolution equations. Brezis and Pazy [11] showed that the solution of such problems is given by nonlinear semigroups whose theory was founded by Crandall, Liggett and Pazy [17], [18]. By using the approximation theorem of Brezis and Pazy, we show the N-stability of the abstract nonlinear discrete problem for the implicit Euler method. Motivated by the rational approximation methods for linear semigroups, we propose a more general time discretization method and prove its N-stability as well.

Basic results in nonlinear theory

In this section we summarise the results about the nonlinear theory we will need. Our main reference is the textbook by Ito and Kappel [37]. We note that another good book in this topic is written by Belleni-Morante and McBride [9].

Let $(\mathscr{X}, \|\cdot\|_{\mathscr{X}})$ denote a Banach space. From now on we identify the operator $A : \mathrm{dom}(A) \subset \mathscr{X} \to \mathscr{X}$ with its graph in $\mathscr{X} \times \mathscr{X}$.

Definition 2.4.3 (Prop. 1.8, [37]). *For $\omega \in \mathbb{R}$, an operator A on $\mathscr{X}$ is called ω-dissipative if for all $\tau \in \left(0, \frac{1}{|\omega|}\right)$ and $f, g \in \mathrm{dom}(A)$ we have*

$$\|(I - \tau A)(f) - (I - \tau A)(g)\|_{\mathscr{X}} \geq (1 - \tau\omega)\|f - g\|_{\mathscr{X}}. \tag{2.66}$$

For $\omega = 0$ the operator A is called dissipative. We note that for $\omega = 0$, we have $\tau \in (0, \infty)$.

Remark 2.4.5. In the literature accretive operators are often considered. We say that operator A is accretive if operator $-A$ in Definition 2.4.3 is dissipative. One of the first papers in this topic is related to Browder [13].

Remark 2.4.6 (Prop. 1.9, [37]). Let A be an ω-dissipative operator on $\mathscr{X}$. Then, for any $\tau \in \left(0, \frac{1}{|\omega|}\right)$, the operator $(I - \tau A)^{-1}$ is single-valued and for any $\tau \in \left(0, \frac{1}{|\omega|}\right)$ and $f, g \in \mathrm{ran}(I - \tau A)$, we have

$$\left\|(I - \tau A)^{-1}(f) - (I - \tau A)^{-1}(g)\right\|_{\mathscr{X}} \leq \frac{1}{1 - \tau\omega}\|f - g\|_{\mathscr{X}}.$$

Definition 2.4.4 (Def. 5.1, [37]). *Let $\mathscr{X}_0$ be a subset of $\mathscr{X}$, $\omega \in \mathbb{R}$ and $(S(t))_{t\geq 0}$ be a family of (nonlinear) operators $\mathscr{X}_0 \to \mathscr{X}_0$. The family $(S(t))_{t\geq 0}$ is called a* strongly continuous semigroup of type ω on $\mathscr{X}_0$ *if it possesses the following properties.*

(i) $S(0)(f) = f$ *for all* $f \in \mathscr{X}_0$.

(ii) $S(t+s)(f) = S(t)S(s)(f)$ *for all* $t, s \geq 0$ *and* $f \in \mathscr{X}_0$.

(iii) For any $f \in \mathscr{X}_0$ *the function* $(0,\infty) \ni t \to S(t)(f) \in \mathscr{X}_0$ *is continuous.*

(iv) There exists $\omega \in \mathbb{R}$ *such that* $\|S(t)(f) - S(t)(g)\|_{\mathscr{X}} \leq \mathrm{e}^{\omega t}\|f-g\|_{\mathscr{X}}$ *for all* $t \geq 0$ *and* $f, g \in \mathscr{X}_0$.

The next celebrated result of Crandall and Liggett shows how one can construct a semigroup by having an appropriate operator at hand.

Theorem 2.4.3 (Thm. I., [17], Cor. 5.4, [37]). *For $\omega \in \mathbb{R}$ let A be an ω-dissipative operator on $\mathscr{X}$ such that* $\mathrm{ran}(I - \tau A) \supset \overline{\mathrm{dom}(A)}$ *holds for every* $\tau \in \left(0, \frac{1}{|\omega|}\right)$. *Then there exists a strongly continuous semigroup $((S(t))_{t\geq 0}$ of type ω on $\overline{\mathrm{dom}(A)}$. Moreover, for $f \in \overline{\mathrm{dom}(A)}$, we have the limit*

$$S(t)(f) = \lim_{k\to\infty} \left(\left(I - \tfrac{t}{k}A\right)^{-1}\right)^k (f) \tag{2.67}$$

which converges uniformly for t in bounded intervals.

In case of the above theorem above we say that the operator A generates the semigroup S. We note that the kth power denotes the k times composition. Next we introduce the relevant results concerning the connection between semigroups of type ω and abstract Cauchy problems with ω-dissipative operators. For an operator A on $\mathscr{X}$ we consider the abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}u(t,\cdot) = A(u(t,\cdot)), \quad t > 0 \\ u(0,\cdot) = u_0(\cdot) \in \mathscr{X}_0. \end{cases} \tag{2.68}$$

For the definition of integral and strong solutions, needed for the next theorem, we refer to Definition 5.5 in Ito and Kappel [37].

Theorem 2.4.4 (Thm. 5.6 and Thm. 5.8, [37]). *Suppose that A is an ω-dissipative opetator on $\mathscr{X}$ generating the strongly continuous semigroup S of type ω. Suppose further that* $\mathrm{ran}(I - \tau A) \supset \overline{\mathrm{dom}(A)}$ *holds for all* $\tau \in \left(0, \frac{1}{|\omega|}\right)$. *Then the following is true.*

i, For any $u_0 \in \overline{\mathrm{dom}(A)}$, there exists a unique integral solution u to problem (2.68) given by $u(t,\cdot) = (S(t)(u_0))(\cdot)$ for all $t \geq 0$.

ii, For $\omega = 0$, the solution above is the unique strong solution.

Later, when studying the convergence of the spatial discretizations, we will need the following theorem as well (similar to Thm. 3.2, [32] and to Cor. 10.8, [37]).

Theorem 2.4.5 (Cor. 4.2, [11]). *Let $\omega \geq 0$ and A be an ω-dissipative single-valued operator on $\mathscr{X}$ satisfying $\overline{\operatorname{ran}(I-\tau A)} \supset \overline{\operatorname{dom}(A)}$ for some $\tau \in \left(0, \frac{1}{\omega}\right)$ and let S be the semigroup of type ω generated by A on $\overline{\operatorname{dom}(A)}$. Let $A_m : \operatorname{dom}(A_m) \subset \mathscr{X} \to \mathscr{X}$ be ω_m-dissipative single-valued operators on $\mathscr{X}$ satisfying $\operatorname{ran}(I-\tau A_m) \supset \overline{\operatorname{dom}(A_m)}$ for some $\tau \in \left(0, \frac{1}{\omega}\right)$ and for all $m \in \mathbb{N}$, and let $\left(S_m(t)\right)_{t\geq 0}$ be the semigroup of type ω_m generated by A_m on $\mathscr{X}$. If*

i, there exists $\alpha \in [0,\infty)$ such that $0 \leq \omega, \omega_m \leq \alpha$,

ii, $\operatorname{dom}(A) \subset \operatorname{dom}(A_m)$ for all $m \in \mathbb{N}$,

iii, $\lim\limits_{m\to\infty} A_m(f) \to A(f)$ for all $f \in \overline{\operatorname{dom}(A)}$,

then we have the limit

$$\lim_{m\to\infty} S_m(t)(f) = S(t)(f) \quad \text{for all} \quad f \in \overline{\operatorname{dom}(A)}, \tag{2.69}$$

where the convergence is uniform for t in bounded intervals.

Discretization schemes

To define the discrete problem (1.4), we consider problem (1.1) with an operator F of a special form. Throughout this section we suppose that A is an ω-dissipative operator on $\mathscr{X}$ for some $\omega \geq 0$ with $\overline{\operatorname{ran}(I-\tau A)} \supset \overline{\operatorname{dom}(A)}$ for some $\tau \in \left(0, \frac{1}{\omega}\right)$. We consider then problem (1.1) in the following form:

$$\begin{cases} F(u) = 0 \quad \text{for} \quad u \in \operatorname{dom}(F), \\ u(0,\cdot) = u_0 \in \operatorname{dom}(A) \quad \text{given}, \\ \big(F(v)\big)(t,x) := \big(\frac{\mathrm{d}}{\mathrm{d}t} v(t,\cdot) - A(v(t,\cdot))\big)(x) \quad \text{for} \quad v \in \operatorname{dom}(F),\ t>0,\ x \in \mathbb{R}^d. \end{cases} \tag{2.70}$$

According to Theorem 2.4.5 operator A generates a semigroup S of type ω on $\overline{\operatorname{dom}(A)}$. In order to obtain an approximation to the exact solution u, i.e., to the semigroup S, we discretise the nonlinear evolution equation (2.70) both in space and time.

Discretization in space

To obtain the spatially discretised solution we assume the following.

Assumption 2.4.2. We assume that there exist operators A_m, $m \in \mathbb{N}$ on $\mathscr{X}$ such that

(a) A_m is ω_m-dissipative on $\mathscr{X}$ for some $\omega_m \geq 0$ for each $m \in \mathbb{N}$,

(b) $\operatorname{ran}(I-\tau A_m) \supset \overline{\operatorname{dom}(A_m)}$ for all $m \in \mathbb{N}$ and for some $\tau \in \left(0, \frac{1}{\omega}\right)$,

(c) there exists $\alpha \in [0,\infty)$ such that $0 \leq \omega, \omega_m \leq \alpha$ for all $m \in \mathbb{N}$,

(d) $\operatorname{dom}(A) \subset \operatorname{dom}(A_m)$ for all $m \in \mathbb{N}$,

(e) $\lim_{m\to\infty} A_m(f) = A(f)$ for all $f \in \overline{\mathrm{dom}(A)}$.

The smallest possible value of α is denoted by β.

Assumption 2.4.2 and Theorem 2.4.3 imply that the operator A_m is the generator of a semigroup S_m for all $m \in \mathbb{N}$. Theorem 2.4.5 implies that these semigroups converge, that is, the limit (2.69) holds uniformly for t in compact intervals.

From the numerical point of view this means that A_m represents the approximation of A by using some spatial discretization scheme. For instance, if A involves a spatial derivative, then A_m stands for e.g. the finite difference approximation or the approximation by using finite discrete Fourier transform. In these cases m refers to the number of spatial grid points or the number of Fourier coefficients, respectively. If the approximate generators A_m converge to A, then the numerical solution will converge to the exact solution, too.

Discretization in time - Implicit Euler method

In order to get the fully discretised approximative solution to problem (1.1) we need to define problem (1.4), especially the operator F_n in it.

First we notice that Theorem 2.4.3 states that the solution u to problem (1.1) has the form $u(t,\cdot) = (S(t)(u_0))(\cdot)$ where S is the semigroup generated by A. Formula (2.67) and Theorem 2.4.5 imply that

$$S(t)(u_0) = \lim_{m\to\infty}\lim_{k\to\infty}\left(\left(I - \tfrac{t}{k}A_m\right)^{-1}\right)^k(u_0), \tag{2.71}$$

where the convergence is uniform for t in compact intervals. We note that limit (2.67) in Theorem 2.4.3 and therefore formula (2.71) already contain a kind of time discretization, namely, the implicit Euler method, that is, when the operator $S_m(t)$ is approximated by the operator $\left(\left(I - \frac{t}{k}A_m\right)^{-1}\right)^k$ for some $k \in \mathbb{N}$. For each $t \geq 0$ we fix now $K \in \mathbb{N}$ such that $K > \beta t$, where β is the smallest possible common bound on ω and ω_m from Assumption 2.4.2 and introduce the product spaces $X_n := \mathscr{X}^{K+1}$, $Y_n := \mathscr{X}^{K+1}$ endowed by some appropriate norms specified later. Then limit (2.71) motivates us how to define the fully discretised numerical solution u_n for all $n \in \mathbb{I}$. Its k^{th} component corresponds to the approximation of the solution at the k^{th} time level, and has the form

$$(u_n)_k = \left(\left(I - \tfrac{t}{K}A_m\right)^{-1}\right)^k(u_0) = \left(I - \tfrac{t}{K}A_m\right)^{-1}\left((u_n)_{k-1}\right) \quad \text{for} \quad k = 0, ..., K. \tag{2.72}$$

Hence, with time step $\tau := \frac{t}{K}$, problem (1.4) contains the operator F_n defined for all $v_n \in (\mathrm{dom}(A))^{K+1}$, $n \in \mathbb{I}$, as

$$\begin{cases} \left(F_n(v_n)\right)_0 := (v_n)_0, \\ \left(F_n(v_n)\right)_k := (v_n)_k - \left(I - \tau A_m\right)^{-1}\left((v_n)_{k-1}\right), \quad \text{for all} \quad k = 1, ..., K, \end{cases} \tag{2.73}$$

where $(v_n)_k \in \mathrm{dom}(A)$ for all $k = 0, ..., K$. Since $\omega_m \leq \beta$ for all $m \in \mathbb{N}$, Remark 2.4.6 implies that for all $f, g \in \mathrm{dom}(A)$ and $m \in \mathbb{N}$ we have

$$\left\|(I - \tau A_m)^{-1}(f) - (I - \tau A_m)^{-1}(g)\right\|_{\mathscr{X}} \leq \Lambda_1 \|f - g\|_{\mathscr{X}}$$
$$\text{with} \quad \Lambda_1 := \frac{1}{1 - \tau\beta}.$$

We note that for dissipative operators A_m we have $\omega_m = 0$, therefore, $\beta = 0$ and $\Lambda_1 = 1$.

Discretization in time - Rational approximations

As we already mentioned rational approximations are well-known and widely investigated for linear operators, see Hairer and Wanner [35]. This motivated us to analyse them in an abstract framework for nonlinear operators as well. For a given $t \geq 0$ we choose $K \in \mathbb{N}$, fix $\tau = \frac{t}{K}$ and choose constants $z_0, z_{ij} \in \mathbb{R}$, $c_i \in \mathbb{R}$, $\nu, \nu_i \in \mathbb{N}$ with $c_i > \beta\tau$ (i.e. $c_i K > \beta t$). Then for all $f \in \mathrm{dom}(A)$ we define the rational approximations for nonlinear operators as

$$r(\tau A_m)(f) = z_0 f + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} z_{ij} \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^j (f). \tag{2.74}$$

After replacing the term $(I - \tau A_m)^{-1}$ by $r(\tau A_m)$ in (2.72), we obtain the discrete problem

$$(u_n)_k = r(\tau A_m)^k (u_0) \quad \text{for} \quad k = 0, ..., K. \tag{2.75}$$

Due to Remark 2.4.6, the operators $(I - \frac{\tau}{c_i} A_m)^{-1} : \mathrm{dom}(A) \to \mathrm{dom}(A)$ exist for all $0 < \frac{\tau}{c_i} < \frac{1}{\beta} < \frac{1}{\omega_m}$, therefore, the operators $r(\tau A_m) : \mathrm{dom}(A) \to \mathrm{dom}(A)$ are well-defined for all $m \in \mathbb{N}$. Formulae (2.73) and (2.75) lead to the full discretization scheme (1.4) with the operator F_n defined for all $v_n \in (\mathrm{dom}(A))^{K+1}$ as

$$\begin{cases} \big(F_n(v_n)\big)_0 = (v_n)_0, \\ \big(F_n(v_n)\big)_k = (v_n)_k - r(\tau A_m)^k \big((v_n)_0\big) \quad \text{for} \quad k = 1, ..., K. \end{cases} \tag{2.76}$$

Remark 2.4.6 implies that for all $f, g \in \mathrm{dom}(A)$ and $m \in \mathbb{N}$ we have

$$\begin{aligned} &\big\|(I - \tfrac{\tau}{c_i} A_m)^{-1}(f) - (I - \tfrac{\tau}{c_i} A_m)^{-1}(g)\big\|_{\mathscr{X}} \leq \Lambda_{c_i} \|f - g\|_{\mathscr{X}} \\ &\text{with} \quad \Lambda_{c_i} := \frac{1}{1 - \frac{\tau}{c_i}\beta}. \end{aligned} \tag{2.77}$$

Hence, for all $f, g \in \mathrm{dom}(A)$ and $m \in \mathbb{N}$ we have the estimate

$$\begin{aligned} &\|r(\tau A_m) f - r(\tau A_m) g\|_{\mathscr{X}} \\ \leq &|z_0| \|f - g\|_{\mathscr{X}} + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \big\| \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^j (f) - \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^j (g) \big\|_{\mathscr{X}} \\ \leq &|z_0| \|f - g\|_{\mathscr{X}} + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i} \big\| \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^{j-1} (f) - \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^{j-1} (g) \big\|_{\mathscr{X}} \\ \leq &|z_0| \|f - g\|_{\mathscr{X}} + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i}^2 \big\| \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^{j-2} (f) - \big((I - \tfrac{\tau}{c_i} A_m)^{-1}\big)^{j-2} (g) \|_{\mathscr{X}} \\ \leq &|z_0| \|f - g\|_{\mathscr{X}} + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i}^j \|f - g\|_{\mathscr{X}} = \Big(|z_0| + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i}^j \Big) \|f - g\|_{\mathscr{X}}. \end{aligned}$$

Thus, by introducing

$$Z := |z_0| + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i}^j \tag{2.78}$$

we have that

$$\|r(\tau A_m)(f) - r(\tau A_m)(g)\|_{\mathscr{X}} \leq Z \|f - g\|_{\mathscr{X}}, \tag{2.79}$$

where Z depends on the choice of the constants in (2.74).

Remark 2.4.7. Since we will use it later, we show now that $Z \geq 1$ holds for the rational approximations defined in (2.74). First we note that the operator $r(\tau A_m)$ is meant to approximate the operator $S_m(\tau)$ which approximates the operator $S(\tau)$. Hence, we expect that $r(\tau A_m)$ should possess some of the properties of $S(\tau)$, one of them is $S(0) = I$. Therefore, it seems natural to expect that $r(0A) = I$ should hold. Then we have that the operator

$$r(0A_m) = z_0 I + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} z_{ij} I$$

equals the identity operator on $\mathscr{X}$ if and only if

$$z_0 + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} z_{ij} = 1.$$

Then the triangular inequality implies that

$$1 \leq |z_0| + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}|.$$

Since $c_i K > \beta t$, from (2.77) we have $\Lambda_{c_i} \geq 1$ for all $\beta \geq 0$, therefore we obtain that

$$Z = |z_0| + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \Lambda_{c_i}^j \geq |z_0| + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} |z_{ij}| \geq 1.$$

At the end of this section we present two basic examples, both being well-known for linear problems, for nonlinear rational approximations (2.74).

Example 2.4.3.

i, The choice $z_0 = 0$, $\nu = 1$, $\nu_1 = 1$, $c_1 = 1$ and $z_{11} = 1$ in (2.74) corresponds to the implicit Euler method with $r(\tau A_m) = (I - \tau A_m)^{-1}$. In case of implicit Euler the estimate (2.79) holds with $Z = \Lambda_1$.

ii, The choice $z_0 = -1$, $\nu = 1$, $\nu_1 = 1$, $c_1 = 2$ and $z_{11} = 2$ gives the Crank-Nicolson method with $r(\tau A_m) = (I + \frac{\tau}{2} A_m)(I - \frac{\tau}{2} A_m)^{-1}$, since by using the identity $(I + \frac{\tau}{2} A_m)(I - \frac{\tau}{2} A_m)^{-1} = I$ we have

$$\begin{aligned} r(\tau A_m) &= -I + 2\big(I - \tfrac{\tau}{2} A_m\big)^{-1} = \big(I - \tfrac{\tau}{2} A_m\big)^{-1} + \big(I - \tfrac{\tau}{2} A_m\big)^{-1} - I \\ &= \big(I - \tfrac{\tau}{2} A_m\big)^{-1} + \tfrac{\tau}{2} A_m \big(I - \tfrac{\tau}{2} A_m\big)^{-1} = \big(I + \tfrac{\tau}{2} A_m\big)\big(I - \tfrac{\tau}{2} A_m\big)^{-1}. \end{aligned}$$

♣

Stability in the nonlinear case

In this section we show the N-stability of the numerical scheme (1.4), that is, $F_n(u_n)$ for $u_n \in \mathrm{dom}(F_n) \subset X_n$, where F_n is defined in (2.76). First we endow the spaces $X_n = \mathscr{X}^{K+1}$ and $Y_n = \mathscr{X}^{K+1}$ by the following norms:

$$\begin{aligned} \|f\|_{X_n} &:= a_K \sum_{k=0}^{K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, ..., f_K) \in X_n = \mathscr{X}^{K+1}, \\ \|f\|_{Y_n} &:= \sum_{k=0}^{K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, ..., f_K) \in Y_n = \mathscr{X}^{K+1}, \end{aligned} \tag{2.80}$$

where

$$a_K = \begin{cases} \dfrac{1}{K+1}, & \text{if } Z = 1, \\ \dfrac{Z-1}{Z^{K+1}-1}, & \text{if } Z > 1. \end{cases} \tag{2.81}$$

Now we are in the position to show N-stability property (2.2) of the general rational approximation schemes defined in (2.76).

Theorem 2.4.6. *Suppose that A is an ω-dissipative operator on $\mathscr{X}$ for some $\omega \geq 0$. Suppose further that the operators A_m, $m \in \mathbb{N}$ satisfy Assumption 2.4.2. Then the numerical scheme (2.76) is N-stable with the stability constant $C = 1$.*

Proof. Since operators A_m are ω_m-dissipative on $\mathscr{X}$ for all $m \in \mathbb{N}$, formula (2.79) implies that

$$\left\| r(\tau A_m)(f) - r(\tau A_m)(g) \right\|_{\mathscr{X}} \leq Z \|f - g\|_{\mathscr{X}}$$

for all $f, g \in \mathrm{dom}(A)$ and $m \in \mathbb{N}$, where Z is defined in (2.78). We have for all $v_n, z_n \in (\mathrm{dom}(A))^{K+1}$ that

$$\begin{aligned}
\|(v_n)_0 - (z_n)_0\|_{\mathscr{X}} &= \left\| \big(F_n(v_n)\big)_0 - \big(F_n(z_n)\big)_0 \right\|_{\mathscr{X}}, \\
\|(v_n)_1 - (z_n)_1\|_{\mathscr{X}} &\leq \left\| \big(F_n(v_n)\big)_1 - \big(F_n(z_n)\big)_1 \right\|_{\mathscr{X}} \\
&\quad + \left\| r(\tau A_m)\big((v_n)_0\big) - r(\tau A_m)\big((z_n)_0\big) \right\|_{\mathscr{X}} \\
&\leq \left\| \big(F_n(v_n)\big)_1 - \big(F_n(z_n)\big)_1 \right\|_{\mathscr{X}} + Z \left\| (v_n)_0 - (z_n)_0 \right\|_{\mathscr{X}} \\
&= \left\| \big(F_n(v_n)\big)_1 - \big(F_n(z_n)\big)_1 \right\|_{\mathscr{X}} + Z \left\| \big(F_n(v_n)\big)_0 - \big(F_n(z_n)\big)_0 \right\|_{\mathscr{X}}, \\
\|(v_n)_2 - (z_n)_2\|_{\mathscr{X}} &\leq \left\| \big(F_n(v_n)\big)_2 - \big(F_n(z_n)\big)_2 \right\|_{\mathscr{X}} \\
&\quad + \left\| r(\tau A_m)\big((v_n)_1\big) - r(\tau A_m)\big((z_n)_1\big) \right\|_{\mathscr{X}} \\
&\leq \left\| \big(F_n(v_n)\big)_2 - \big(F_n(z_n)\big)_2 \right\|_{\mathscr{X}} + Z \left\| (v_n)_1 - (z_n)_1 \right\|_{\mathscr{X}} \\
&= \left\| \big(F_n(v_n)\big)_2 - \big(F_n(z_n)\big)_2 \right\|_{\mathscr{X}} + Z \left\| \big(F_n(v_n)\big)_1 - \big(F_n(z_n)\big)_1 \right\|_{\mathscr{X}} \\
&\quad + Z^2 \left\| \big(F_n(v_n)\big)_0 - \big(F_n(z_n)\big)_0 \right\|_{\mathscr{X}}.
\end{aligned}$$

Therefore, there exists an index $\ell \in \mathbb{N}$ such that

$$\|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \leq \sum_{j=0}^{k} Z^{k-j} \left\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \right\|_{\mathscr{X}} \tag{2.82}$$

holds for all $k = 0, ..., \ell$. The definition (2.76) of F_n and the estimate (2.82) yields

$$\begin{aligned}
&\|(v_n)_{\ell+1} - (z_n)_{\ell+1}\|_{\mathscr{X}} \\
\leq &\left\| \big(F_n(v_n)\big)_{\ell+1} - \big(F_n(z_n)\big)_{\ell+1} \right\|_{\mathscr{X}} + \left\| r(\tau A_m)\big((v_n)_\ell\big) - r(\tau A_m)\big((z_n)_\ell\big) \right\|_{\mathscr{X}} \\
\leq &\left\| \big(F_n(v_n)\big)_{\ell+1} - \big(F_n(z_n)\big)_{\ell+1} \right\|_{\mathscr{X}} + Z \|(v_n)_\ell - (z_n)_\ell\|_{\mathscr{X}} \\
\leq &\left\| \big(F_n(v_n)\big)_{\ell+1} - \big(F_n(z_n)\big)_{\ell+1} \right\|_{\mathscr{X}} + Z \sum_{j=0}^{\ell} Z^{\ell-j} \left\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \right\|_{\mathscr{X}} \\
= &\sum_{j=0}^{\ell+1} Z^{\ell+1-j} \left\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \right\|_{\mathscr{X}}.
\end{aligned}$$

By induction we obtain that (2.82) holds for all $k \in \mathbb{N}$, which we repeat here for further references:

$$\|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \leq \sum_{j=0}^{k} Z^{k-j} \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \quad \text{for all} \quad k \in \mathbb{N}. \tag{2.83}$$

From this point we have two cases: $Z = 1$ and $Z > 1$.

The case $Z = 1$. Estimate (2.83) has now the form

$$\|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \leq \sum_{j=0}^{k} \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \quad \text{for all} \quad k \in \mathbb{N}. \tag{2.84}$$

Inserting (2.84) into the definition (2.80) of the norm leads to the estimate

$$\begin{aligned}
\|v_n - z_n\|_{X_n} &= \frac{1}{K+1} \sum_{k=0}^{K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\leq \frac{1}{K+1} \sum_{k=0}^{K} \sum_{j=0}^{k} \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \\
&= \frac{1}{K+1} \sum_{j=0}^{K} (K+1-j) \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \qquad (2.85) \\
&\leq \frac{1}{K+1} \sum_{k=0}^{K} (K+1) \big\| \big(F_n(v_n)\big)_k - \big(F_n(z_n)\big)_k \big\|_{\mathscr{X}} \\
&= \sum_{k=0}^{K} \big\| \big(F_n(v_n)\big)_k - \big(F_n(z_n)\big)_k \big\|_{\mathscr{X}} = \|F_n(v_n) - F_n(z_n)\|_{Y_n}.
\end{aligned}$$

This yields N-stability with $C = 1$.

The case $Z > 1$. From formula (2.83) we obtain the estimate

$$\|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \leq \sum_{j=0}^{k} Z^{k-j} \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \quad \text{for all} \quad k \in \mathbb{N}. \tag{2.86}$$

In the same manner as before, we insert (2.86) into the definition (2.80) and obtain

$$\begin{aligned}
\|v_n - z_n\|_{X_n} &= \frac{Z-1}{Z^{K+1}-1} \sum_{k=0}^{K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\leq \frac{Z-1}{Z^{K+1}-1} \sum_{k=0}^{K} \sum_{j=0}^{k} Z^{k-j} \big\| \big(F_n(v_n)\big)_j - \big(F_n(z_n)\big)_j \big\|_{\mathscr{X}} \\
&= \frac{Z-1}{Z^{K+1}-1} \sum_{k=0}^{K} \frac{Z^{K+1-k}-1}{Z-1} \big\| \big(F_n(v_n)\big)_k - \big(F_n(z_n)\big)_k \big\|_{\mathscr{X}} \qquad (2.87) \\
&\leq \frac{Z-1}{Z^{K+1}-1} \sum_{k=0}^{K} \frac{Z^{K+1}-1}{Z-1} \big\| \big(F_n(v_n)\big)_k - \big(F_n(z_n)\big)_k \big\|_{\mathscr{X}} \\
&= \sum_{k=0}^{K} \big\| \big(F_n(v_n)\big)_k - \big(F_n(z_n)\big)_k \big\|_{\mathscr{X}} = \|F_n(v_n) - F_n(z_n)\|_{Y_n}.
\end{aligned}$$

This yields N-stability with $C = 1$ in this case as well. ∎

We note that the case $Z = 1$ corresponds e.g. to the implicit Euler method for dissipative operators ($\beta = 0$).

Remark 2.4.8. We briefly show that Theorem 2.4.6 remains valid if the norms are defined different from (2.80).

(a) We endow the spaces $X_n = \mathscr{X}^{K+1}$ and $Y_n = \mathscr{X}^{K+1}$ with the following norms:

$$\|f\|_{X_n} := a_K \sup_{k=0,\dots,K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, \dots, f_K) \in X_n,$$
$$\|f\|_{Y_n} := \sup_{k=0,\dots,K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, \dots, f_K) \in Y_n,$$

where a_K is defined as before in (2.81) and $f_k \in \mathscr{X}$ for all $k = 0, \dots, K$. The proof of Theorem 2.4.6 has to be changed only at the last estimates (2.85) and (2.87), respectively.

Estimate (2.83) implies for $Z = 1$ that

$$\begin{aligned}
\|v_n - z_n\|_{X_n} &= \frac{1}{K+1} \sup_{k=0,\dots,K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\le \frac{1}{K+1} \sup_{k=0,\dots,K} \left(\sum_{j=0}^{k} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \right) \\
&\le \frac{1}{K+1} \sup_{k=0,\dots,K} \left(\sum_{j=0}^{k} \sup_{j=0,\dots,K} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \right) \\
&\le \frac{1}{K+1} \sup_{k=0,\dots,K} (k+1) \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n} \\
&= \frac{1}{K+1} (K+1) \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n} = \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n}.
\end{aligned}$$

This yields N-stability with $C = 1$ indeed.

Estimate (2.83) implies for $Z > 1$ that

$$\begin{aligned}
&\|v_n - z_n\|_{X_n} \\
&= \frac{Z-1}{Z^{K+1}-1} \sup_{k=0,\dots,K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\le \frac{Z-1}{Z^{K+1}-1} \sup_{k=0,\dots,K} \left(\sum_{j=0}^{k} Z^{k-j} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \right) \\
&\le \frac{Z-1}{Z^{K+1}-1} \left(\sup_{k=0,\dots,K} \sum_{j=0}^{k} Z^j \right) \left(\sup_{j=0,\dots,K} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \right) \\
&\le \frac{Z-1}{Z^{K+1}-1} \left(\sup_{k=0,\dots,K} \frac{Z^{k+1}-1}{Z-1} \right) \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n} \\
&= \frac{Z-1}{Z^{K+1}-1} \frac{Z^{K+1}-1}{Z-1} \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n} = \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n}.
\end{aligned}$$

Which yields N-stability again with $C = 1$.

(b) Now we endow the spaces $X_n = \mathscr{X}^{K+1}$ and $Y_n = \mathscr{X}^{K+1}$ with the following norms:

$$\|f\|_{X_n} := \frac{1}{Z^K} \sup_{k=0,\dots,K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, \dots, f_K) \in X_n = \mathscr{X}^{K+1},$$

$$\|f\|_{Y_n} := \sum_{k=0}^{K} \|f_k\|_{\mathscr{X}} \quad \text{for} \quad f = (f_0, \dots, f_K) \in Y_n = \mathscr{X}^{K+1}.$$

Using (2.83) for $Z = 1$ one obtains now the following estimate instead of (2.85) in the proof of Theorem 2.4.6:

$$\begin{aligned}
\|v_n - z_n\|_{X_n} &= \sup_{k=0,\dots,K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\le \sup_{k=0,\dots,K} \left(\sum_{j=0}^{k} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \right) \\
&= \sum_{j=0}^{K} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} = \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n}
\end{aligned}$$

meaning again N-stability with $C = 1$.

Using (2.83) we have for $Z > 1$ that

$$\begin{aligned}
\|v_n - z_n\|_{X_n} &= \frac{1}{Z^K} \sup_{k=0,\dots,K} \|(v_n)_k - (z_n)_k\|_{\mathscr{X}} \\
&\le \frac{1}{Z^K} \sup_{k=0,\dots,K} \sum_{j=0}^{k} Z^{k-j} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \\
&\le \frac{1}{Z^K} Z^K \sup_{k=0,\dots,K} \sum_{j=0}^{k} \left\| \left(F_n(v_n)\right)_j - \left(F_n(z_n)\right)_j \right\|_{\mathscr{X}} \\
&\le \frac{1}{Z^K} Z^K \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n} = \left\| F_n(v_n) - F_n(z_n) \right\|_{Y_n}.
\end{aligned}$$

Which yields again N-stability with $C = 1$.

Stability in the linear case

In this section we show how our results apply for the linear case. We take the same setting (spaces and norms) as defined by Sanz-Serna and Palencia in [54] for linear operators. Our aim is to show that Example 3.1 in [54], that is, the classical Lax–Richtmyer theory, follows from our recent results for rational approximations defined by formula (2.74).

Let $F : \mathrm{dom}(F) \subset X \to Y$ be the operator defined in (2.70), where A is now a linear operator on the Banach space $\mathscr{X}$. As in [54], let the spaces $X_n = \mathscr{X}^{K+1}$, $Y_n = \mathscr{X}^{K+1}$ be endowed by the norms

$$\|f\|_{X_n} = \sup_{k=0,\dots,K} \|f_k\|_{\mathscr{X}} \quad \text{for all} \quad f \in X_n \quad \text{and}$$

$$\|f\|_{Y_n} = \sum_{k=0}^{K} \|f_k\|_{\mathscr{X}} \quad \text{for all} \quad f \in Y_n,$$

respectively, that is, the case (b) in Remark 2.4.8 without the multiplication by a_K. Then condition (2.2) of the N-stability of the operator $F_n : \mathrm{dom}(F_n) \subset X_n \to Y_n$, reduces to the estimate $\|F_n^{-1}\|_{Y_n \to X_n} \leq \widetilde{C}$ for some constant $\widetilde{C} > 0$. Let F_n be defined as in (2.76) with the linear operators A_m, $m \in \mathbb{N}$, satisfying Assumptions 2.4.2, and the rational approximation r defined in (2.74). In this case we have

$$F_n = \begin{pmatrix} I & 0 & 0 & \dots & 0 \\ -r(\tau A_m) & I & 0 & \dots & 0 \\ 0 & -r(\tau A_m) & I & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -r(\tau A_m) & I \end{pmatrix}$$

and

$$F_n^{-1} = \begin{pmatrix} I & 0 & 0 & 0 & \dots & 0 \\ r(\tau A_m) & I & 0 & 0 & \dots & 0 \\ r(\tau A_m)^2 & r(\tau A_m) & I & 0 & \dots & 0 \\ r(\tau A_m)^3 & r(\tau A_m)^2 & r(\tau A_m) & I & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ r(\tau A_m)^K & r(\tau A_m)^{K-1} & r(\tau A_m)^{K-2} & \dots & r(\tau A_m) & I \end{pmatrix},$$

which are exactly the same operator matrices presented in [54]. The norm of F_n^{-1} can be estimated as

$$\begin{aligned}
\|F_n^{-1}\|_{Y_n \to X_n} &= \sup_{\substack{f \in Y_n \\ \|f\|_{X_n}=1}} \|F_n^{-1} f\|_{X_n} \\
&= \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \left\| \begin{pmatrix} f_0 \\ r(\tau A_m) f_0 + f_1 \\ r(\tau A_m)^2 f_0 + r(\tau A_m) f_1 + f_2 \\ \vdots \\ r(\tau A_m)^K f_0 + r(\tau A_m)^{K-1} f_1 + \dots + r(\tau A_m) f_{K-1} + f_K \end{pmatrix} \right\|_{X_n} \\
&= \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \sup_{k=0,\dots,K} \left\| \sum_{j=0}^{k} r(\tau A_m)^j f_{k-j} \right\|_{\mathscr{X}} \\
&\leq \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \sup_{k=0,\dots,K} \sum_{j=0}^{k} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}} \|f_{k-j}\|_{\mathscr{X}} \\
&\leq \sup_{j=0,\dots,K} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}} \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \sup_{k=0,\dots,K} \sum_{j=0}^{k} \|f_{k-j}\|_{\mathscr{X}} \\
&\leq \sup_{j=0,\dots,K} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}} \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \sup_{k=0,\dots,K} \sum_{j=0}^{k} \|f_j\|_{\mathscr{X}} \\
&\leq \sup_{j=0,\dots,K} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}} \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \sum_{j=0}^{K} \|f_j\|_{\mathscr{X}} \\
&= \sup_{j=0,\dots,K} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}} \sup_{\substack{f \in Y_n \\ \|f\|_{Y_n}=1}} \|f\|_{Y_n} = \sup_{j=0,\dots,K} \|r(\tau A_m)^j\|_{\mathscr{X} \to \mathscr{X}}.
\end{aligned}$$

Hence, one obtains the following stability condition: There should exists a contant $\widetilde{C} > 0$ such that

$$\sup_{k=0,\dots,K} \|r(\tau A_m)^k\|_{\mathscr{X}\to\mathscr{X}} \leq \widetilde{C} \tag{2.88}$$

holds for all $\tau = \frac{t}{K}$ for each fixed $t \geq 0$ time level. For a fixed $K \in \mathbb{N}$, this is the usual definition of Lax–Richtmyer stability obtained in [54] as well. Since formula (2.79) corresponds to $\|r(\tau A_m)\|_{\mathscr{X}\to\mathscr{X}} \leq Z$ for linear operators, we have that

$$\sup_{k=0,\dots,K} \|r(\tau A_m)^k\|_{\mathscr{X}\to\mathscr{X}} \leq \sup_{k=0,\dots,K} \|r(\tau A_m)\|^k_{\mathscr{X}\to\mathscr{X}} \leq \sup_{k=0,\dots,K} Z^k = Z^K, \tag{2.89}$$

that is, in this case the stability criterion (2.88) holds with $\widetilde{C} := Z^K$ for each fixed $K \in \mathbb{N}$. We note that if the norms are defined as in Remark 2.4.8/(b), then (2.89) is the same result as stated in Theorem 2.4.6.

CHAPTER 3

Other stability notions

3.1 Necessity of N-stability

In Theorem 2.0.1 we have shown that in case of consistency N-stability is sufficient to guarantee convergence. However, its necessity is not clear. In this section we investigate this question. Using an example taken from [44], we will show that the N-stability requirement is too restrictive.

Let $F_n^\alpha : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$ be the operator given as

$$[F_n^\alpha(\mathbf{z})]_k = \begin{cases} \dfrac{z_k - z_{k-1}}{\tau} - z_{k-1}^2, & k = 1, 2, \ldots, K, \\ \\ z_0 - \alpha, & k = 0, \end{cases} \tag{3.1}$$

where τ is the step-size parameter, $\alpha \in [0, 1)$ is some fixed constant and $K\tau = 1$. Taking the function $\bar{z}^\alpha(t) = \alpha/[1 - \alpha t]$, where $t \in [0, 1]$ and applying φ_n as a grid function to the function $\bar{z}^\alpha(t)$, we get

$$[\varphi_n(\bar{z}^\alpha)]_k \equiv (\bar{\mathbf{z}}_n^\alpha)_k \equiv \bar{z}^\alpha(t_k) \equiv \frac{\alpha}{1 - \alpha t_k}, \quad k = 0, 1, \ldots, K,$$

where t_k are the grid points.

Remark 3.1.1. With the discrete operator (3.1) the problem $F_n^\alpha(u_n) = 0$ can be considered as the discretization of the prototype of the simple Ricatti equation:

$$\begin{cases} u'(t) = u^2(t), \ t \in [0, 1], \\ \\ u(0) = \alpha, \end{cases} \tag{3.2}$$

by means of the explicit Euler's rule on the equidistant mesh. Clearly, the solution of the problem (3.2) is the function $\bar{z}^\alpha$.

Substituting $\bar{\mathbf{z}}_n^\alpha$ into (3.1), we gain

$$[F_n^\alpha(\bar{\mathbf{z}}_n^\alpha)]_k = \begin{cases} \dfrac{\bar{z}^\alpha(t_k) - \bar{z}^\alpha(t_{k-1})}{\tau} - [\bar{z}^\alpha(t_{k-1})]^2, & k = 1, 2, \ldots, K, \\ \\ (\bar{z}_n^\alpha)_0 - \alpha, & k = 0. \end{cases}$$

Let $\bar{\mathbf{w}}_n \in \mathbb{R}^{K+1}$ be a vector with the components w_k, such that $[F_n(\bar{\mathbf{w}}_n)] = 0$, where

$$[F_n(\bar{\mathbf{w}}_n)]_k = \begin{cases} \dfrac{w_k - w_{k-1}}{\tau} - w_{k-1}^2, & k = 1, 2, \dots, K, \\ w_0 - 1, & k = 0. \end{cases}$$

We introduce the norms

$$\|\mathbf{x}_k\|_{X_n} = \max_{1 \le k \le K+1} |x_k|,$$

$$\|\mathbf{y}_k\|_{Y_n} = |y_0| + \sum_{k=1}^{K} h|y_k|,$$

respectively. We prove that the stability estimate (2.2) cannot be true for any stability constant C independent of the mesh size. To this aim, we show that the estimate

$$\|\bar{\mathbf{z}}_n^\alpha - \bar{\mathbf{w}}_n\|_{X_n} \le C \, \|F_n^\alpha(\bar{\mathbf{z}}_n^\alpha) - F_n(\bar{\mathbf{w}}_n)\|_{Y_n} \tag{3.3}$$

cannot hold uniformly for all n.

Since $(\bar{w}_n)$ is defined by the recursion $\bar{\mathbf{w}}_n = \bar{\mathbf{w}}_{n-1} + \tau \bar{\mathbf{w}}_n^2$, due to [55], the approximation at the last grid point $t = 1$ behaves like $1/(\tau|\ln \tau|)$. Thus,

$$\lim_{K \to \infty} (\bar{\mathbf{w}}_n)_K = \lim_{\tau \to 0} \frac{1}{\tau |\ln \tau|} = \infty.$$

Since $(\bar{\mathbf{z}}_n^\alpha)_K \equiv \alpha/[1-\alpha]$ and $\alpha \in [0, 1)$, the value of $(\bar{\mathbf{z}}_n^\alpha)_K$ is finite. So the left term of (3.3) converges to ∞ as $n \to \infty$, i.e.

$$\lim_{n \to \infty} \|\bar{\mathbf{z}}_n^\alpha - \bar{\mathbf{w}}_n\|_{X_n} = \infty. \tag{3.4}$$

For the right-hand side of (3.3) we have

$$[F_n^\alpha(\bar{\mathbf{z}}_n^\alpha) - \underbrace{F_n(\bar{\mathbf{w}}_n)}_{=0}]_k = \begin{cases} \dfrac{\bar{z}^\alpha(t_k) - \bar{z}^\alpha(t_{k-1})}{\tau} - [\bar{z}^\alpha(t_{k-1})]^2, & k = 1, 2, \dots, K \\ \alpha - 1, & k = 0. \end{cases} \tag{3.5}$$

Using the introduced norm in Y_n to (3.5) and Lemma A.2.4, we get

$$\|F_n^\alpha(\bar{\mathbf{z}}_n^\alpha) - F_n(\bar{\mathbf{w}}_n)\|_{Y_n} = |\alpha - 1| + \sum_{k=1}^{K} \tau \cdot l_n(\bar{z}^\alpha(t_k)) \le |\alpha - 1| + \frac{M_2(\bar{z}^\alpha)}{2}.$$

Thus,

$$\lim_{n \to \infty} \|F_n^\alpha(\bar{\mathbf{z}}_n^\alpha) - F_n(\bar{\mathbf{w}}_n)\|_{Y_n} < \infty. \tag{3.6}$$

From (3.4) and (3.6) we can see the estimate (3.3) cannot hold. This means that the discretization is not N-stable.

Thus, the statement of Theorem 2.0.1 cannot be satisfied. However, we will see through the numerical results that the forward Euler method on the equidistant mesh will converge to the solution of the problem (3.2). To demonstrate this, we select the value $\alpha = 0.8$ in (3.2), and we apply the forward Euler method to this problem. The results have been summarized in Figure 3.1 and Table 3.1. The obtained numerical results suggest the convergence of the method.

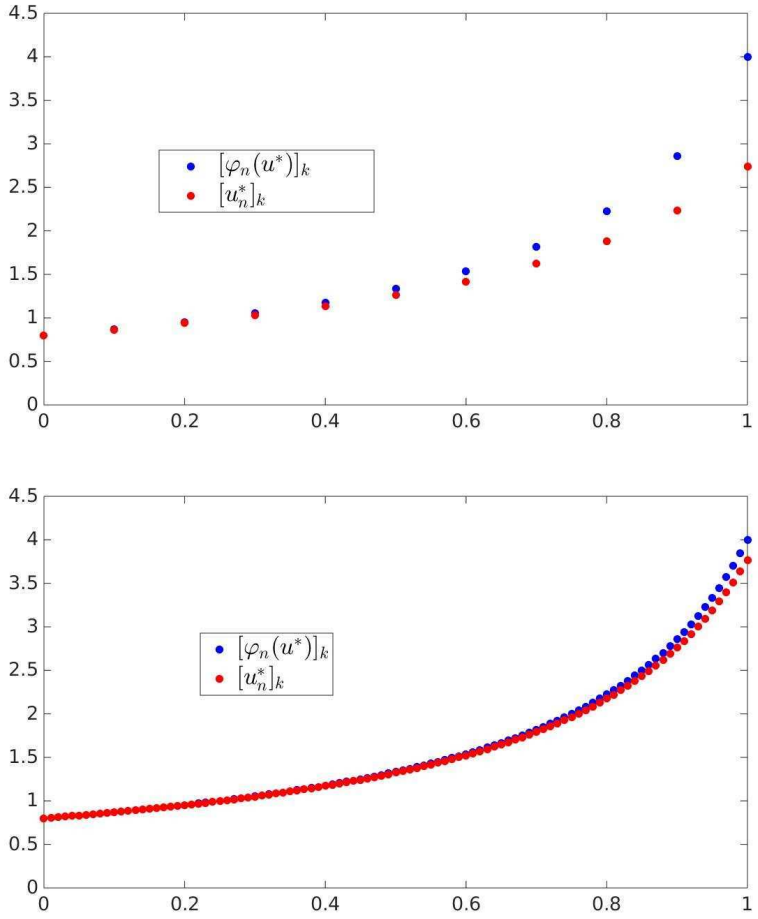


Figure 3.1. *The restricted true solution and the numerical solution for* 10 *and* 100 *grid points to the problem* (3.2).

Number of grid points	$\|e_n\|_{X_n}$
10^1	$1.5175 \cdot 10^1$
10^2	$5.8687 \cdot 10^{-1}$
10^4	$6.0863 \cdot 10^{-2}$
10^6	$6.0887 \cdot 10^{-3}$

Table 3.1. *The global discretization error in the introduced norm to the problem* (3.2).

3.2 K-stability

Section 3.1 shows that the N-stability definition is too restrictive, because we require the condition (2.2) for any elements from $\mathrm{dom}(F_n)$. It also shows that if $\bar{\mathbf{w}}_n$ is far from $\bar{\mathbf{z}}_n^\alpha$ (i.e., the perturbation $\bar{\mathbf{z}}_n^\alpha$ is too large), then the estimate (2.2) cannot hold.

3.2.1 Theoretical results

This motivates to introduce the idea of local stability and stability threshold notions [40]. The first step in this direction is done by introducing a simplified form of the notion of semistability in [44].

Definition 3.2.1. *The discretization $\mathscr{D}$ is called semistable on problem $\mathscr{P}$ if there exist $C \in \mathbb{R}$, $R \in (0, \infty]$ such that*

i, $B_R(\varphi_n(u^*)) \subset \mathrm{dom}(F_n)$ *holds from some index,*

ii, $\forall (v_n)_{n\in\mathbb{I}}$ *which satisfy* $v_n \in B_R(\varphi_n(u^*))$ *from that index, the relation*

$$\|\varphi_n(u^*) - v_n\|_{X_n} \leq C \, \|F_n(\varphi_n(u^*)) - F_n(v_n)\|_{Y_n}$$

holds.

Semistability is a purely theoretical notion, which, similarly to the consistency, cannot be checked directly, due to the fact that u^* is unknown. However, the following statement clearly shows the relation of the three important notions.

Lemma 3.2.1. *We assume that*

i, Assumption 1.1.1 (a) is fulfilled,

ii, discretization $\mathscr{D}$ is consistent in order p at u^* and semistable with stability threshold R on problem $\mathscr{P}$,

iii, discretization $\mathscr{D}$ generates a numerical method $\mathscr{N}$ that equation (1.4) has solution u_n^* in $B_R(\varphi_n(u^*))$ from some index.

Then the sequence of these solutions of (1.4) converges to the solution of (1.1), and the order of convergence is not less than the order of consistency.

Proof. Using i, and iii, we have the relation $F_n(u_n^*) = 0$ and $\psi_n(F(\bar{u})) = 0$, respectively. Therefore, we gain for the global discretization error that

$$\|\varphi_n(u^*) - u_n^*\|_{X_n} \leq C \, \|F_n(\varphi_n(u^*)) - F_n(u_n^*)\|_{Y_n} = C\|l_n\|_{Y_n},$$

which proves the statement. ∎

This lemma has some drawbacks. First, we cannot verify its conditions because this requires the knowledge of the solution. Secondly, we have no guarantee that equation (1.4) has a (possibly unique) solution in $B_R(\varphi_n(u^*))$ from some index. The following modified stability notion related to Keller (see [40]) gets rid of the second problem.

Definition 3.2.2. *The discretization $\mathscr{D}$ is called* K-stable *for problem $\mathscr{P}$ on the element $v \in \mathcal{X}$ if there exist a positive constant C and $R \in (0, \infty]$ such that*

i, $B_R(\varphi_n(v)) \subset \mathrm{dom}(F_n)$ *holds from some index,*

ii, for all $z_n,\ w_n \in B_R(\varphi_n(v))$ *the estimate*

$$\|z_n - w_n\|_{X_n} \leq C \|F_n(z_n) - F_n(w_n)\|_{Y_n} \tag{3.7}$$

holds.

Remark 3.2.1. Obviously, Definition 3.2.2 implies Definition 3.2.1.

The immediate profit of this definition is injectivity as it is formulated in the next statement.

Corollary 3.2.2. *If discretization $\mathscr{D}$ is K-stable on problem $\mathscr{P}$ on the element $v \in X$ with stability threshold R, then F_n is injective on $B_R(\varphi_n(v))$ from some index.*

The following lemma shows the importance of the K-stability notion. The proof based on Stetter's train of thought (see [56]).

Lemma 3.2.3. *We assume that*

i, V, W *are normed spaces with the property* $\dim V = \dim W < \infty$,

ii, $G : B_R(v) \to W$ *is continuous, where* $B_R(v) \subset V$ *is a ball for some* $v \in V$ *and* $R \in (0, \infty]$,

iii, for all v^1, v^2 *which satisfy* $v^i \in B_R(v)$, $i = 1, 2$ *the estimate*

$$\left\|v^1 - v^2\right\|_V \leq C \left\|G(v^1) - G(v^2)\right\|_W \tag{3.8}$$

holds.

Then

i, G *is invertible, and* $G^{-1} : B_{R/C}(G(v)) \to B_R(v)$;

ii, G^{-1} *is Lipschitz continuous with the constant* C.

Proof. Due to Corolarry 3.2.2 it is enough to show that $B_{R/C}(G(v)) \subset G(B_R(v))$. Assuming indirectly that there exists $w \in B_{R/C}(G(v))$ such that $w \notin G(B_R(v))$. Furthermore, we define the line $w(\lambda) = (1 - \lambda)G(v) + \lambda w$ for $\lambda \geq 0$ and introduce the number $\hat{\lambda}$ as follows:

$$\hat{\lambda} := \begin{cases} \sup\left\{\lambda' > 0 \mid w(\lambda) \in G(B_R(v)) \text{ for all } \lambda \in [0, \lambda')\right\}, \text{ if it exists,} \\ 0, \text{ else.} \end{cases}$$

In this case the inequality $\hat{\lambda} \leq 1$ holds. We will show that $\hat{w} := w(\hat{\lambda}) \in G(B_R(v))$.

- For $\hat{\lambda} = 0$ this trivially holds.
- For $\hat{\lambda} > 0$ we observe that G is invertible on $w(\hat{\lambda} - \varepsilon)$ for all $\varepsilon \in (0, \hat{\lambda}]$. It means that the operators $G^{-1}(w(\hat{\lambda} - \varepsilon)) \in B_R(v)$ exist. Hence, we can use stability estimate (3.8)

$$\begin{aligned} \left\|G^{-1}(w(\hat{\lambda} - \varepsilon)) - v\right\|_V &\leq C \left\|w(\hat{\lambda} - \varepsilon) - G(v)\right\|_W \\ &= C(\hat{\lambda} - \varepsilon) \left\|w - G(v)\right\|_W \\ &< \hat{\lambda}(R - \delta) \\ &\leq R - \delta \end{aligned}$$

for some $\delta > 0$. Using again the stability estimate we can conclude that the function $h(\varepsilon) = G^{-1}(w(\hat{\lambda} - \varepsilon))$ is uniformly continuous at $\varepsilon \in (0, \hat{\lambda}]$. Hence, there exists $\lim\limits_{\varepsilon \searrow 0} h(\varepsilon)$. It is denoted by z and we know that $z \in B_R(v)$. Using the continuity of G, we get $G(z) = \hat{w}$. Due to Brouwer's invariance domain theorem (see [12]) we can choose a closed ball $\bar{B}_r(z) \subset B_R(v)$ whose image $G(\bar{B}_r(z))$ contains a neighbourhood of $\hat{w}$. This results in a contradiction.

Finally, the Lipschitz continuity with the constant C is a simple consequence of estimate 3.8. ■

Assumption 3.2.1. The operator F_n is continuous on the ball $B_R(\varphi_n(u^*))$.

Lemma 3.2.4. *Assume that*

i, discretization $\mathscr{D}$ is consistent and K-stable at u^* with stability threshold R and constant C on problem $\mathscr{P}$,

ii, Assumptions 1.1.1 and 3.2.1 are fulfilled.

Then discretization $\mathscr{D}$ generates a numerical method $\mathscr{N}$ such that equation (1.4) has a unique solution in $B_R(\varphi_n(u^*))$ from some index.

Proof. Due to Lemma 3.2.3 the operator F_n is invertible and we also know that $F_n^{-1} : B_{R/C}(F_n(\varphi_n(u^*))) \to B_R(\varphi_n(u^*))$. Due to Assumption 1.1.1 the consistency is $F_n(\varphi_n(u^*)) = l_n \to 0$. This means that $0 \in B_{R/C}(F_n(\varphi_n(u^*)))$ from some index. This proves the statement. ■

Hence, we can formulate the main result of this section.

Theorem 3.2.5. *Assume that*

i, discretization $\mathscr{D}$ is consistent in order p and K-stable at u^* with stability threshold R and constant C on problem $\mathscr{P}$,

ii, Assumptions 1.1.1 and 3.2.1 are fulfilled.

Then discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$ and the order of convergence is not less than the order of consistency.

Proof. The statement is a simple consequence of Lemmas 3.2.1 and 3.2.4. ■

Remark 3.2.2. Lemma 3.2.4 guarantees that equation (1.4) has a unique solution in some suitably chosen ball. This means that K-stability in the nonlinear case locally satisfies those properties what the linear stability notion (or, equivalently, the N-stability notion for the linear case) does.

3.2.2 K-stability for a general class of operators

Let $F_n : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$ be the operator given as

$$[F_n(\mathbf{z})]_k = \begin{cases} \dfrac{z_k - z_{k-1}}{\tau} - f(z_{k-1}), \; k = 1, 2, \ldots, K \\ \\ z_0 - u_0, \qquad\qquad\qquad k = 0, \end{cases} \tag{3.9}$$

where τ is the step-size parameter, $f : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz continuous function and u_0 is some fixed value. The discretization (3.9) is the application of the explicit Euler method on the equidistant mesh to the autonomous Cauchy problem

$$\begin{cases} u'(t) = f(u(t)), \; t \in [0, 1], \\ \\ u(0) = u_0. \end{cases} \tag{3.10}$$

Let $R > 0$ and $B_R = \bigcup\limits_{t\in[0,1]} [u(t) - R, u(t) + R]$. The function f is Lipschitz continuous on B_R with constant $L(R)$. We consider only those vectors $\mathbf{z}_n$, $\mathbf{w}_n$ for which

$$\|\mathbf{z}_n - \varphi_n(u^*)\|_{X_n} \leq R$$

and

$$\|\mathbf{w}_n - \varphi_n(u^*)\|_{X_n} \leq R.$$

These conditions imply that $(\mathbf{z}_n)_k$, $(\mathbf{w}_n)_k \in B_R$, where the Lipschitz condition holds. Then we substitute $\mathbf{z}_n$ and $\mathbf{w}_n$ into (3.9). The subtraction of $[F_n(\mathbf{z}_n)]_k$ and $[F_n(\mathbf{w}_n)]_k$ leads to the equality

$$\begin{aligned}(\mathbf{z}_n)_k - (\mathbf{w}_n)_k = (\mathbf{z}_n)_{k-1} - (\mathbf{w}_n)_{k-1} + \tau\Big([f(\mathbf{z}_n)]_{k-1} - [f(\mathbf{w}_n)]_{k-1}\Big) \\ + \tau\Big([F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k\Big).\end{aligned}$$

Using the Lipschitz condition we gain

$$|(\mathbf{z}_n)_k - (\mathbf{w}_n)_k| \leq (1 + \tau L(R))|(\mathbf{z}_n)_{k-1} - (\mathbf{w}_n)_{k-1}| + \tau|[F_n(\mathbf{z}_n)]_k - [F_n(\mathbf{w}_n)]_k|.$$

Then, by induction we get

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} = \max_{0\leq k\leq K} |(\mathbf{z}_n)_k - (\mathbf{w}_n)_k| \leq e^{L(R)} \|F_n(\mathbf{z}_n) - F_n(\mathbf{w}_n)\|_{Y_n}. \tag{3.11}$$

The estimate (3.11) is in the form of (3.7), i.e. the discretization - which is consistent - is K-stable with stability constant $C = e^{L(R)}$.

Theorem 3.2.6. *The discrete operator (3.9) under the given conditions is K-stable with the stability constant $C = e^{L(R)}$.*

Hence, in virtue of Theorems 3.2.5 and 3.2.6 the following statement is true.

Corollary 3.2.7. *The sequence of the solutions of the problems $F_n(\boldsymbol{z}_n) = 0$ (where F_n is defined by (3.9)) is convergent to the solution of the Cauchy problem (3.10).*

Remark 3.2.3. We recall the discretization (3.1) and the problem (3.2). As we have seen in Section 3.1, the discretization is not N-stable. However, if we choose $f(u(t)) \equiv u^2(t)$ and $u_0 \equiv \alpha \in [0, 1)$ in Theorem 3.2.6, one can see that the discretization is K-stable.

Remark 3.2.4. Let $R > 0$ fixed. Then, as we have seen in Section 3.1, the condition $\bar{\mathbf{v}}_n^\alpha, \bar{\mathbf{w}}_n \in B_R(\bar{\mathbf{v}}_n^\alpha)$ cannot be guaranteed. However, if we require the stability condition only for the elements from $B_R(\bar{\mathbf{v}}_n^\alpha)$ (that is the stability notion in Definition 3.2.2 as we have seen in the previous example for a general class of operators), then the condition (3.7) is satisfied.

In a similar way we examine K-stability for a more general class of discrete operators. Let $F_n^\theta : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ be the operator given as

$$[F_n^\theta(\mathbf{z})]_k = \begin{cases} \dfrac{z_k - z_{k-1}}{\tau} - (1-\theta) f(z_{k-1}) - \theta f(z_k), & k = 1, 2, \ldots, K \\ z_0 - u_0, & k = 0, \end{cases} \tag{3.12}$$

where $\theta \in [0, 1]$ is a given parameter, τ denotes the step-size, $f : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz continuous function and u_0 is some fixed value. The discretization (3.9) can be viewed as the application of the standard θ-method on the equidistant mesh to the problem (3.10).

In the previous train of thought we get the equality

$$\begin{aligned}(\mathbf{z}_n)_k - (\mathbf{w}_n)_k = (\mathbf{z}_n)_{k-1} - (\mathbf{w}_n)_{k-1} + \tau(1-\theta)\Big([f(\mathbf{z}_n)]_{k-1} - [f(\mathbf{w}_n)]_{k-1}\Big) \\ + \tau\theta\Big([f(\mathbf{z}_n)]_k - [f(\mathbf{w}_n)]_k\Big) + \tau\Big([F_n^\theta(\mathbf{z}_n)]_k - [F_n^\theta(\mathbf{w}_n)]_k\Big).\end{aligned}$$

Using the Lipschitz condition we obtain

$$\begin{aligned}\Big|(\mathbf{z}_n)_k - (\mathbf{w}_n)_k\Big| \le \frac{1 + \tau(1-\theta)L(R)}{1 - \tau\theta L(R)}\Big|(\mathbf{z}_n)_{k-1} - (\mathbf{w}_n)_{k-1}\Big| \\ + \frac{1}{1 - \tau\theta L(R)}\tau\Big|[F_n^\theta(\mathbf{z}_n)]_k - [F_n^\theta(\mathbf{w}_n)]_k\Big|.\end{aligned}$$

Hereinafter, based on [23], we give an estimation for $(1 - \tau\theta L(R))^{-1}$. For the values τ, satisfying the condition $\tau\theta L(R) \in [0, 1/2]$, we have

$$1 \le \frac{1}{1 - \tau\theta L(R)} = 1 + \tau\theta L(R) + (\tau\theta L(R))^2 \frac{1}{1 - \tau\theta L(R)}.$$

Hence, the estimate

$$\frac{(\tau\theta L(R))^2}{1 - \tau\theta L(R)} \le \tau\theta L(R)$$

holds. Therefore, we have the upper bound

$$\frac{1}{1 - \tau\theta L(R)} \le 1 + 2\tau\theta L(R).$$

Thus, we can give the following estimate

$$\Big|(\mathbf{z}_n)_k - (\mathbf{w}_n)_k\Big| \le (1+2\tau\theta L(R))\Big[(1+\tau(1-\theta)L(R))\Big|(\mathbf{z}_n)_{k-1} - (\mathbf{w}_n)_{k-1}\Big| + \tau\Big|[F_n^\theta(\mathbf{z}_n)]_k - [F_n^\theta(\mathbf{w}_n)]_k\Big|\Big].$$

Then, by induction we get

$$\|\mathbf{z}_n - \mathbf{w}_n\|_{X_n} = \max_{0\le k\le K} |(\mathbf{z}_n)_k - (\mathbf{w}_n)_k| \le e^{(1+\theta)L(R)}\|F_n^\theta(\mathbf{z}_n) - F_n^\theta(\mathbf{w}_n)\|_{Y_n}. \quad (3.13)$$

The estimate (3.13) proves the validity of the following statement.

Theorem 3.2.8. *The discrete operator* (3.12) *is K-stable with the stability constant* $C = e^{(1+\theta)L(R)}$.

Due to consistency, in virtue of Theorems 3.2.5 and 3.2.8, the following statement is true.

Corollary 3.2.9. *The sequence of the solutions of the problems* $F_n^\theta(\boldsymbol{z}_n) = 0$ *(where* F_n^θ *is defined by* (3.12)*) is convergent to the solution of the Cauchy problem* (3.10).

3.3 T-stability

In the 1980's V. A. Trenogin laid down the foundations of this topic in [62]. Namely, by giving the definition of T-stability, the explicit Euler method is considered on an equidistant grid and its T-stablility for the initial-value problem is proven. First we consider Trenogin's stability definition.

Definition 3.3.1. *The discretization* $\mathscr{D}$ *is called T-stable if there exists a continuous, strictly monotonically increasing function* $\omega(s)$*, defined for* $s \ge 0$*, such that* $\omega(0) = 0$ *and* $\omega(\infty) = \infty$ *and*

$$\omega\left(\|z_n - w_n\|_{X_n}\right) \le \|F_n(z_n) - F_n(w_n)\|_{Y_n} \quad (3.14)$$

holds for all $z_n, w_n \in \mathrm{dom}(F_n)$.

Several theoretical results derived from Definition 3.3.1 can be found in [62]. Here we add our further results. In this section we assume that the norm consistency, i.e. relation (1.7) in Remark 1.2.1 is fulfilled. Furthermore, we also assume that $\varphi_n \in B(X, X_n)$.

In case of relation (1.7), i.e. consistent norms the following property is valid.

Lemma 3.3.1. *When the norms* $\|\cdot\|_{X_n}$ *are consistent to the norm* $\|\cdot\|_X$*, then the relation* $v = 0$ *is valid if and only if* $\lim\limits_{n\to\infty} \|\varphi_n(v)\|_{X_n} = 0$.

Proof. We consider two cases.

i, If $v = 0$, then $\lim\limits_{n\to\infty} \|\varphi_n(v)\|_{X_n} = \|v\|_X = 0$.

ii, If $\lim\limits_{n\to\infty} \|\varphi_n(v)\|_{X_n} = 0$, then $\|v\|_X = 0$. Hence, $v = 0$.

This proves the statement. ∎

Generally, when $\lim_{n\to\infty} \|\varphi_n(v)\|_{X_n} = 0$ if and only if $v = 0$, we say that the spaces X_n are regularly normed. Hence, when $\|\cdot\|_{X_n}$ is consistent to the norm $\|\cdot\|_X$, then X_n are regularly normed spaces.

Theorem 3.3.2. *Suppose that*

i, the sequence of norms $\|\cdot\|_{\mathcal{X}_n}$ is consistent to the norm $\|\cdot\|_{\mathcal{X}}$,

ii, there exists a solution to the problems (1.1) and (1.4),

iii, discretization $\mathscr{D}$ is consistent and T-stable at the element u^*.

Then

i, u^* is unique,

ii, for any $n \in \mathbb{I}$ the discrete solution u_n^* is unique,

iii, the numerical method $\mathscr{N}$ is convergent.

Proof.

i, Let v_1, v_2 be solutions of (1.1) and assume that for these elements the relations

$$\lim_{n\to\infty} \|F_n(\varphi_n(v_1))\|_{Y_n} = 0 \text{ and } \lim_{n\to\infty} \|F_n(\varphi_n(v_2))\|_{Y_n} = 0$$

hold. Then we gain

$$\begin{aligned}\|\varphi_n(v_1 - v_2)\|_{X_n} &\le \omega^{-1}(\|F_n(\varphi_n(v_1)) - F_n(\varphi_n(v_2))\|)_{Y_n} \\ &\le \omega^{-1}(\|F_n(\varphi_n(v_1))\|_{Y_n} + \|F_n(\varphi_n(v_2))\|_{Y_n}) \to 0,\end{aligned}$$

if n tends to ∞. Hence, we get

$$\lim_{n\to\infty} \|\varphi_n(v_1 - v_2)\|_{X_n} = 0.$$

Since the finite dimensional spaces X_n are regularly normed, the solution is unique.

ii, Let v_n^1 and v_n^2 be two solutions of (1.4). Substituting into (3.14), we get

$$0 = \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{Y_n} \ge \omega\left(\left\|v_n^1 - v_n^2\right\|_{X_n}\right) \ge 0.$$

It means that $\omega\left(\|v_n^1 - v_n^2\|_{X_n}\right) = 0$. From the norm property it follows that $v_n^1 = v_n^2$.

iii, Let v_1 and v_n^1 be solutions of (1.1) and (1.4), respectively. From the Definition 3.3.1 we gain

$$\left\|v_n^1 - \varphi(v_1)\right\|_{X_n} \le \omega^{-1}(\left\|F_n(v_n^1) - F_n(\varphi(v_1))\right\|)_{Y_n} = \omega^{-1}(\|F_n(\varphi(v_1))\|)_{Y_n},$$

where we have used the consistency and also the continuity of the function ω^{-1} at the point $t = 0$, i.e. it approaches zero when n tends to ∞.

This proves the statement. ∎

3.3.1 T-stability of one-step methods for the initial-value problem

In this part we revise Definition 3.3.1 from the application point of view. We consider the initial-value problem from Example 1.1.1. In the sense of Definitions 1.1.1, 1.1.2 and 1.1.3 the operators F, φ_n, ψ_n and the spaces X, Y, X_n, Y_n are defined in Examples 1.1.1, 1.1.2 and 1.1.3, respectively.

In the following section we will define precisely the operators F_n and Φ_n for the explicit and implicit one-step methods. To verify the T-stability of a given method applied to problem (1.2)-(1.3), we consider the equation

$$F_n(\mathbf{x}_n + \mathbf{z}_n) - F_n(\mathbf{x}_n) = \mathbf{y}_n, \tag{3.15}$$

where $\mathbf{x}_n$ is some parameter and $\mathbf{z}_n$ is unknown. If we can give an estimation in the form of

$$\|\mathbf{z}_n\|_{X_n} \leq \zeta(\|\mathbf{y}_n\|_{Y_n}), \tag{3.16}$$

where the properties of $\zeta(s)$ correspond to the properties of $\omega(s)$, then by the choice $\omega(s) := \zeta^{-1}(s)$ we prove T-stability.

Let $\mathbf{x}_n = \mathbf{x}_n^1$, while $\mathbf{x}_n + \mathbf{z}_n = \mathbf{x}_n^2$ in (3.15). Then $F_n(\mathbf{x}_n^2) - F_n(\mathbf{x}_n^1) = \mathbf{y}_n$ and $\mathbf{x}_n^2 - \mathbf{x}_n^1 = \mathbf{z}_n$. Based on estimation (3.16), we get

$$\left\|\mathbf{x}_n^2 - \mathbf{x}_n^1\right\|_{X_n} \leq \zeta(\left\|F_n(\mathbf{x}_n^2) - F_n(\mathbf{x}_n^1)\right\|_{Y_n}).$$

Because the inverse of ζ exists and it is strictly monotonically increasing, we have

$$\zeta^{-1}\Big(\left\|\mathbf{x}_n^2 - \mathbf{x}_n^1\right\|_{X_n}\Big) \leq \left\|F_n(\mathbf{x}_n^2) - F_n(\mathbf{x}_n^1)\right\|_{Y_n}.$$

This matches the stability estimate in Definition 3.3.1. Hence, in order to verify T-stability, we have to prove that the estimate

$$\|\mathbf{z}_n\|_{X_n} \leq c\,\|\mathbf{y}_n\|_{Y_n} = c\Big(\max_{1\leq k\leq K}|y_k| + |u_0|\Big) \tag{3.17}$$

holds.

Our goal is to generalize Trenogin's result. He showed that under a natural assumption the explicit Euler method is T-stable for (1.2)-(1.3) on an equidistant grid. In the following section we will prove that any explicit or implicit one-step method is T-stable for (1.2)-(1.3) on a non-equidistant grid, too.

First of all we define the non-equidistant grid as

$$\mathbb{G}_n := \{\tau_k = t_k - t_{k-1},\ k = 1, \ldots, K,\ 0 = t_0 < t_1 < \ldots < t_K = T\}. \tag{3.18}$$

The general form of the one-step method can be written as

$$y_k = y_{k-1} + \tau_k \Lambda(t_{k-1}, y_{k-1}, y_k, \tau_k), \tag{3.19}$$

where $\Lambda : \mathbb{R}^4 \to \mathbb{R}$ defines the given the numerical method $\mathscr{N}$. The function Λ is often called the increment function and can be interpreted as an estimate of the slope of y_k.

Remark 3.3.1. We will say that the methods are explicit if $\Lambda = \Lambda(t_{i-1}, y_{i-1}, \tau_k)$. In the other case the methods are implicit.

3.3.2 Explicit one-step methods

In this part we consider the case where the numerical method is explicit. To this aim, we define the operators F_n and Φ_n from Definitions 1.1.2 and 1.1.3, respectively.

$F_n : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$ and for any $\mathbf{v}_n = (v_0, v_1, \dots, v_K) \in \mathbb{R}^{K+1}$ it acts as

$$[F_n(\mathbf{v}_n)]_k = \begin{cases} \dfrac{v_k - v_{k-1}}{\tau_k} - \Lambda\left(t_{k-1}, v_{k-1}, \tau_k\right), & k = 1, \dots, K, \\ v_0 - u_0, & k = 0. \end{cases} \tag{3.20}$$

In order to give Φ_n, we define the mapping $\Phi_n : C^1[0,T] \to \mathbb{R}^{K+1}$ in the following way:

$$[(\Phi_n(F))\,(\varphi_n(u))]_k = \begin{cases} \dfrac{u(t_k) - u(t_{k-1})}{\tau_k} - \Lambda(t_{k-1}, u(t_{k-1}), \tau_k), & k = 1, \dots, K, \\ u(t_0) - u_0, & k = 0. \end{cases} \tag{3.21}$$

In the sequel we assume that Λ is a Lipschitz continuous function with respect to its second variable, by the constant L_Λ. It means there exists a constant $L_\Lambda \geq 0$ such that for arbitrary $s_1, s_2 \in \mathbb{R}$ the estimation

$$|\Lambda(t_{k-1}, s_1, \tau_k) - \Lambda(t_{k-1}, s_2, \tau_k)| \leq L_\Lambda |s_1 - s_2| \tag{3.22}$$

holds for $t_{k-1} \in \mathbb{G}_n$ and $\tau_k > 0$.

Remark 3.3.2. The Lipschitz assumption (3.22) is obviously necessary in proving convergence. For the explicit Runge–Kutta methods this condition can be guaranteed directly: when the Lipschitz assumption for the function f in (1.2) is valid, then the increment function Λ of the eligible explicit Runge–Kutta method satisfies the condition (3.22). For non-autonomous problems Lipschitz condition is the same w.r.t. the second variable.

Substituting F_n into (3.15) and (3.20), we gain

$$\begin{cases} \dfrac{z_k - z_{k-1}}{\tau_k} = y_i + \Lambda(t_{k-1}, x_{k-1} + z_{k-1}, \tau_k) - \Lambda(t_{k-1}, x_{k-1}, \tau_k) & k = 1, \dots, K, \\ z_0 = u_0, & k = 0. \end{cases} \tag{3.23}$$

From (3.23) we get the estimate

$$|z_k| \leq (1 + L_\Lambda \tau_k)|z_{k-1}| + \tau_k |y_k|, \;\; k = 1, \dots, K. \tag{3.24}$$

The equidistant case

For each index k writing out (3.24) and applying it recursively, we get

$$\begin{aligned} |z_1| &\le (1+L_\Lambda\tau)|u_0| + \tau\,\|\mathbf{y}_n\|_\infty\,,\\ |z_2| &\le (1+L_\Lambda\tau)^2|u_0| + \tau\,\|\mathbf{y}_n\|_\infty\,[1+(1+L_\Lambda\tau)],\\ &\vdots\\ |z_K| &\le \underbrace{(1+L_\Lambda\tau)^K}_{(I)}\,|u_0| + \|\mathbf{y}_n\|_\infty\,\underbrace{\tau\sum_{k=0}^{K-1}(1+L_\Lambda\tau)^k}_{(II)}\,. \end{aligned} \tag{3.25}$$

In the next step we estimate the terms on the right-hand side of (3.25).

$$(I) \le e^{L_\Lambda\tau K} = e^{L_\Lambda T}$$

$$(II) \le \tau\sum_{k=0}^{K-1}(1+L_\Lambda\tau)^K = \tau\frac{(1+L_\Lambda\tau)^K-1}{1+L_\Lambda\tau-1} \le \tau\frac{e^{L_\Lambda\tau K}-1}{L_\Lambda\tau} = \frac{e^{L_\Lambda T}-1}{L_\Lambda}$$

Then we get for the norm of $\mathbf{z}_n$ the estimate

$$\|\mathbf{z}_n\|_{X_n} \le e^{L_\Lambda T}|u_0| + \|\mathbf{y}_n\|_\infty\,\frac{e^{L_\Lambda T}-1}{L_\Lambda} \le c\,\|\mathbf{y}_n\|_{Y_n}$$

with the choice

$$c = \max\left(e^{L_\Lambda T}, \frac{e^{L_\Lambda T}-1}{L_\Lambda}\right).$$

This implies the validity of the estimate (3.17). Hence, we have proved the following statement.

Theorem 3.3.3. *Under the condition (3.22) the explicit one-step methods are T-stable for (1.2)-(1.3) on an equidistant grid .*

Remark 3.3.3. For $\Lambda(t_{i-1}, x_{i-1}, h) \equiv f(t_{i-1}, x_{i-1})$ we obtain the explicit Euler method on a equidistant grid. Therefore, Theorem 3.3.3 implies Trenogin's basic result. The constant c is the same which is given in [62].

The non-equidistant case

When the grid is non-equidistant, i.e., the step size is not constant, we can use the previous formula. Namely, for each index k writing out (3.24) and applying it recursively, we get

$$\begin{aligned} |z_1| &\le (1+L_\Lambda\tau_1)|u_0| + \tau_1\,\|\mathbf{y}_n\|_\infty\,,\\ &\vdots\\ |z_K| &\le \underbrace{(1+L_\Lambda\tau_K)\cdot\ldots\cdot(1+L_\Lambda\tau_1)}_{(I)}\,|u_0|\\ &\quad + \|\mathbf{y}_n\|_\infty\,\underbrace{\sum_{k=1}^{K}\tau_k(1+L_\Lambda\tau_{k+1})\cdot\ldots\cdot(1+L_\Lambda\tau_K)}_{(II)}\,. \end{aligned} \tag{3.26}$$

In the next step we estimate the terms on the right-hand side of (3.26).

$$(I) \leq e^{L_\Lambda \tau_n} \cdot \ldots \cdot e^{L_\Lambda \tau_1} = e^{L_\Lambda(\tau_n + \ldots + \tau_1)} = e^{L_\Lambda T}$$

$$(II) \leq \sum_{k=1}^{K} \tau_k e^{(1-t_k)L_\Lambda} = e^{L_\Lambda} \sum_{k=1}^{K} \tau_k (e^{-L_\Lambda})^{t_k} < e^{L_\Lambda} \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} \left(e^{-L_\Lambda}\right)^t \mathrm{d}t$$

$$= e^{L_\Lambda} \int_0^T \left(e^{-L_\Lambda}\right)^t \mathrm{d}t = e^{L_\Lambda} \Big[-\frac{1}{L_\Lambda} e^{-L_\Lambda t} \Big]_0^T = \frac{e^{L_\Lambda} - e^{L_\Lambda(1-T)}}{L_\Lambda}$$

Then for the norm of $\mathbf{z}_n$ we get the estimate

$$\|\mathbf{z}_n\|_{X_n} \leq e^{L_\Lambda T}|u_0| + \|\mathbf{y}_n\|_\infty \frac{e^{L_\Lambda} - e^T}{L_\Lambda} \leq c \|\mathbf{y}_n\|_{Y_n}$$

with the choice

$$c = \max\left(e^{L_\Lambda T}, \frac{e^{L_\Lambda} - e^{L_\Lambda(1-T)}}{L_\Lambda}\right).$$

Therefore, we obtain the estimate in the form (3.17). Hence, like for the equidistant meshes, we have proved the following statement.

Theorem 3.3.4. *Under the condition* (3.22) *explicit one-step methods are T-stable for* (1.2)-(1.3) *on a non-equidistant grid.*

Remark 3.3.4. In case of zero-stability known from linear theory we get similar results to the explicit Euler method, i.e. in case of uniform and non-uniform grid the same constant c can be achieved. Here we come to the same conclusion for T-stability explicit one-step methods.

3.3.3 Implicit one-steps methods

In this part we move on to the consideration of implicit one-step methods. We have to define again the operators F_n and Φ_n.

$F_n : \mathbb{R}^{K+1} \to \mathbb{R}^{K+1}$ and for any $\mathbf{v}_n = (v_0, v_1, \ldots, v_K) \in \mathbb{R}^{K+1}$ it acts as

$$[F_n(\mathbf{v}_n)]_k = \begin{cases} \dfrac{v_k - v_{k-1}}{\tau_k} - \Lambda\left(t_{k-1}, v_{k-1}, v_k, \tau_k\right), \; k = 1, \ldots, K, \\ \\ v_0 - u_0, \qquad\qquad\qquad\qquad k = 0. \end{cases} \tag{3.27}$$

In order to give Φ_n, we define the mapping $\Phi_n : C^1[0,1] \to \mathbb{R}^{n+1}$ in the following way:

$$[(\Phi_n(F)(\varphi_n(u))]_k = \begin{cases} \dfrac{u(t_k) - u(t_{k-1})}{\tau_k} - \Lambda(t_{k-1}, u(t_{k-1}), u(t_k), \tau_k), \; k = 1, \ldots, K, \\ \\ u(t_0) - u_0, \qquad\qquad\qquad\qquad k = 0. \end{cases} \tag{3.28}$$

In the following we suppose that Λ is a Lipschitz continuous function with respect to its second and third variable, by the constants L_{Λ_1} and L_{Λ_2}. It means there exist $L_{\Lambda_1}, L_{\Lambda_2} \geq 0$ constants, such that for arbitrary $s_1, s_2, p_1, p_2 \in \mathbb{R}$ the estimate

$$|\Lambda(t_{k-1}, s_1, p_1, \tau_k) - \Lambda(t_{k-1}, s_2, p_2, \tau_k)| \leq L_{\Lambda_1}|s_1 - s_2| + L_{\Lambda_2}|p_1 - p_2| \tag{3.29}$$

holds for $t_{k-1} \in \mathbb{G}_n$ and $\tau_k > 0$.

Remark 3.3.5. For the implicit Runge–Kutta methods the Lipschitz assumption (3.29) can be also guaranteed directly: when the Lipschitz assumption for the function f in (1.2) is valid, then the increment function Λ of the eligible implicit Runge–Kutta method for a sufficiently small τ_k satisfies the condition (3.29).

Substituting F_n into (3.15), (3.27) and using the Lipschitz condition (3.29), we get the estimate

$$|z_k| \leq |z_{k-1}| + \tau_k|y_k| + \tau_k(L_{\Lambda_1}|z_{k-1}| + L_{\Lambda_2}|z_k|),\ k = 1, \dots, K.$$

Hence, we get

$$|z_k| \leq \frac{1 + \tau_k L_{\Lambda_1}}{1 - \tau_k L_{\Lambda_2}}|z_{k-1}| + \frac{1}{1 - \tau_k L_{\Lambda_2}}\tau_k|y_k|,\ k = 1, \dots, K. \tag{3.30}$$

We give an estimation for $\dfrac{1}{1 - \tau_k L_{\Lambda_2}}$. If $\tau_k L_{\Lambda_2} \in [0, 0.5]$ for all k, then we can write this expression as

$$1 \leq \frac{1}{1 - \tau_k L_{\Lambda_2}} = 1 + \tau_k L_{\Lambda_2} + (\tau_k L_{\Lambda_2})^2 \frac{1}{1 - \tau_k L_{\Lambda_2}}.$$

Obviously, for the values $\tau_k L_{\Lambda_2} \in [0, 0.5]$ the following estimate holds:

$$\frac{(\tau_k L_{\Lambda_2})^2}{1 - \tau_k L_{\Lambda_2}} \leq \tau_k L_{\Lambda_2}.$$

Therefore, we have the upper bound

$$\frac{1}{1 - \tau_k L_{\Lambda_2}} \leq 1 + 2\tau_k L_{\Lambda_2} \leq \exp(2\tau_k L_{\Lambda_2}).$$

Thus, we can write equation (3.30) in the form

$$|z_k| \leq (1 + \tau_k L_{\Lambda_1})(1 + 2\tau_k L_{\Lambda_2})|z_{k-1}| + (1 + 2\tau_k L_{\Lambda_2})\tau_k|y_k|,\ k = 1, \dots, K. \tag{3.31}$$

The equidistant case

For each index k writing out (3.31) and applying it recursively, we get

$$\begin{aligned} |z_1| &\leq (1 + \tau L_{\Lambda_1})(1 + 2\tau L_{\Lambda_2})|u_0| + (1 + 2\tau L_{\Lambda_2})\tau \left\|\mathbf{y}_n\right\|_\infty, \\ &\vdots \\ |z_K| &\leq \underbrace{\Big((1 + \tau L_{\Lambda_1})(1 + 2\tau L_{\Lambda_2})\Big)^K}_{(I)}|u_0| + \left\|\mathbf{y}_n\right\|_\infty \underbrace{\tau \sum_{k=1}^{K}(1 + \tau L_{\Lambda_1})^{k-1}(1 + \tau L_{\Lambda_2})^k}_{(II)}. \end{aligned} \tag{3.32}$$

In the next step we estimate the terms on the right-hand side of (3.32).

$$\begin{aligned} (I) &\leq e^{\tau K L_{\Lambda_1}} e^{2\tau K L_{\Lambda_2}} = e^{L_{\Lambda_1} T} e^{2L_{\Lambda_2} T} = e^{T(L_{\Lambda_1} + 2L_{\Lambda_2})} \\ (II) &\leq \tau \sum_{k=1}^{K} \Big[(1 + \tau L_{\Lambda_1})(1 + \tau L_{\Lambda_2})\Big]^k \leq \tau \frac{\Big[(1 + \tau L_{\Lambda_1})(1 + \tau L_{\Lambda_2})\Big]^{K+1} - 1}{(1 + \tau L_{\Lambda_1})(1 + \tau L_{\Lambda_2}) - 1} \\ &\leq \tau \frac{e^{T(L_{\Lambda_1} + 2L_{\Lambda_2})} - 1}{\tau L_{\Lambda_1} + 2\tau^2 L_{\Lambda_1} L_{\Lambda_2} + 2\tau L_{\Lambda_2}} \leq \frac{e^{T(L_{\Lambda_1} + 2L_{\Lambda_2})} - 1}{L_{\Lambda_1} + 2L_{\Lambda_2}} \end{aligned}$$

Then for the norm of $\mathbf{z}_n$ we get estimate

$$\|\mathbf{z}_n\|_{X_n} \le e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}|u_0| + \|\mathbf{y}_n\|_\infty \frac{e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}-1}{L_{\Lambda_1}+2L_{\Lambda_2}} \le c\,\|\mathbf{y}_n\|_{Y_n}$$

with the choice

$$c = \max\left(e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}, \frac{e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}-1}{L_{\Lambda_1}+2L_{\Lambda_2}}\right).$$

Hence, we obtain the estimate (3.17), which shows the validity of the following statements.

Theorem 3.3.5. *Under the condition* (3.29) *the implicit one-step numerical methods are T-stable for* (1.2)-(1.3) *on an equidistant grid.*

The non-equidistant case

Similarly to the previous case, for each index k writing out (3.31) and applying it recursively, we get

$$\begin{aligned}
|z_1| &\le (1+\tau_1 L_{\Lambda_1})(1+2\tau_1 L_{\Lambda_2})|u_0| + (1+2\tau_1 L_{\Lambda_2})\tau_1\,\|\mathbf{y}_n\|_\infty\,,\\
&\vdots\\
|z_K| &\le \underbrace{(1+\tau_n L_{\Lambda_1})\cdot\ldots\cdot(1+\tau_1 L_{\Lambda_1})(1+2\tau_n L_{\Lambda_2})\cdot\ldots\cdot(1+2\tau_1 L_{\Lambda_2})}_{(I)}|u_0|\\
&\quad + \|\mathbf{y}_n\|_\infty \underbrace{\sum_{k=1}^{K}\tau_k(1+\tau_{k+1}L_{\Lambda_1})\cdot\ldots\cdot(1+\tau_K L_{\Lambda_1})(1+2\tau_k L_{\Lambda_2})\cdot\ldots\cdot(1+2\tau_K L_{\Lambda_2})}_{(II)}
\end{aligned} \tag{3.33}$$

In the next step we estimate the terms on the right-hand side of (3.33).

$$\begin{aligned}
(I) &\le e^{L_{\Lambda_1}(\tau_K+\ldots+\tau_1)}e^{2L_{\Lambda_2}(\tau_K+\ldots+\tau_1)} = e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}\\
(II) &\le \sum_{k=1}^{K}\tau_k e^{(1-t_k)L_{\Lambda_1}}e^{(1-t_{k-1})2L_{\Lambda_2}} = e^{L_{\Lambda_1}+2L_{\Lambda_2}}\sum_{k=1}^{K}\tau_k\left(e^{-L_{\Lambda_1}}\right)^{t_k}\left(e^{-2L_{\Lambda_2}}\right)^{t_{k-1}}\\
&< e^{L_{\Lambda_1}+2L_{\Lambda_2}}\sum_{k=1}^{K}\int_{t_{k-1}}^{t_k}\left(e^{-[L_{\Lambda_1}+2L_{\Lambda_2}]}\right)^t \mathrm{d}t = e^{L_{\Lambda_1}+2L_{\Lambda_2}}\int_0^T\left(e^{-[L_{\Lambda_1}+2L_{\Lambda_2}]}\right)^t \mathrm{d}t\\
&= \frac{e^{L_{\Lambda_1}+2L_{\Lambda_2}} - e^{(L_{\Lambda_1}+2L_{\Lambda_2})(1-T)}}{L_{\Lambda_1}+2L_{\Lambda_2}}
\end{aligned}$$

Then for the norm of $\mathbf{z}_n$ we get the estimate

$$\|\mathbf{z}_n\|_{X_n} \le e^{L_{\Lambda_1}+2L_{\Lambda_2}}|u_0| + \|\mathbf{y}_n\|_\infty \frac{e^{L_{\Lambda_1}+2L_{\Lambda_2}} - e^{(L_{\Lambda_1}+2L_{\Lambda_2})(1-T)}}{L_{\Lambda_1}+2L_{\Lambda_2}} \le c\,\|\mathbf{y}_n\|_{Y_n}$$

with the choice

$$c = \max\left(e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}, \frac{e^{L_{\Lambda_1}+2L_{\Lambda_2}} - e^{(L_{\Lambda_1}+2L_{\Lambda_2})(1-T)}}{L_{\Lambda_1}+2L_{\Lambda_2}}\right).$$

We can formulate our main result in the form of the following statements.

Theorem 3.3.6. *Under the condition (3.29) the implicit one-step methods are T-stable for (1.2)-(1.3) on a non-equidistant grid.*

Table 3.2 summarizes the stability results of Sections 3.3.2 and 3.3.3 for the different cases of the given one-step methods. Due to Theorem 3.3.2 the obtained T-stability together with consistency ensures convergence. It is known that consistency of one-step methods can be given by the following two properties (see, e.g. [41]): the Lipschitz condition and the increment function Λ for the function $f = 0$ should be identically zero, i.e. $\Lambda\left(t_{i-1}, v_{i-1}, v_i, h_i\right) = 0$ for all $t_{k-1} \in \mathbb{G}_n$.

	explicit one-step methods	implicit one-step methods
τ	$\max\left(e^{L_\Lambda T}, \dfrac{e^{L_\Lambda T}-1}{L_\Lambda}\right)$	$\max\left(e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}, \dfrac{e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}-1}{L_{\Lambda_1}+2L_{\Lambda_2}}\right)$
τ_k	$\max\left(e^{L_\Lambda T}, \dfrac{e^{L_\Lambda}-e^{L_\Lambda(1-T)}}{L_\Lambda}\right)$	$\max\left(e^{T(L_{\Lambda_1}+2L_{\Lambda_2})}, \dfrac{e^{L_{\Lambda_1}+2L_{\Lambda_2}}[1-e^{1-T}]}{L_{\Lambda_1}+2L_{\Lambda_2}}\right)$

Table 3.2. *T-stability constants of the different cases.*

3.4 Notes on further stability notions

We finish this section with some remarks with respect to introduced stability notions by the Definitions 2.0.5, 3.2.2 and 3.3.1. There are other definitions of the stability in the literature, these are mostly generalizations of the K-stability notion. Here we will list two of them.

The first one is related to Stetter and it is given in [56].

Definition 3.4.1. *The discretization $\mathscr{D}$ is called S-stable on problem $\mathscr{P}$ if there exist a positive stability constant C, a stability threshold $R \in (0, \infty]$ and $r \in (0, \infty]$ such that*

i, $B_R(\varphi_n(u^*)) \subset \mathrm{dom}(F_n)$ *holds from some index,*

ii, in case of $z_n, w_n \in B_R(\varphi_n(u^*))$ *and* $F_n(z_n), F_n(w_n) \in B_r(F_n(\varphi_n(u^*)))$*, the estimate*

$$\|z_n - w_n\|_{X_n} \leq C \, \|F_n(z_n) - F_n(w_n)\|_{Y_n}$$

holds.

Note that the stability notion by Stetter is less restrictive than the one given in Definition 3.2.2. If we put $r = \infty$ in Definition 3.4.1, then we re-obtain the K-stability notion in 3.2.2. In [56] we can find a similar theorem to Theorem 3.2.5, but this notion seem to be too theoretical.

The last stability notion allows us to vary the radius of the balls which could be necessary as it has been shown in the paper [45] by López-Marcos and Sanz-Serna.

Definition 3.4.2. *The discretization $\mathscr{D}$ is called LSS-stable on problem $\mathscr{P}$ if there exist a positive stability constant C and a varying threshold $R_n \in (0, \infty]$ such that*

i, $B_{R_n}(\varphi_n(u^*)) \subset \mathrm{dom}(F_n)$ *holds from some index,*

ii, for all z_n, w_n *which satisfy* $z_n, w_n \in B_{R_n}(\varphi_n(u^*))$ *from an index, the estimate*

$$\|z_n - w_n\|_{X_n} \leq C \, \|F_n(z_n) - F_n(w_n)\|_{Y_n}$$

holds.

In paper [1] the authors use a similar abstract framework and Definition 3.4.2 to formulate schemes for the numerical solution to a hierarchically size-structured population model. Paper [2] presents an efficient numerical method for the approximation of a nonlinear size-structured population model.

CHAPTER 4

Basic notions revisited

The main result of Section 3.2 is not yet suitable for our purposes, since the condition of Theorem 3.2.5 requires to check the stability and the consistency on the unknown element u^*. Typically we are able to verify the above properties on some set of points (in an ideal case on the entire $\mathrm{dom}(F)$) which includes u^*. Therefore, we extend the previously given pointwise (local) definitions to the set (global) ones.

4.1 Set definitions of the basic notions

Definition 4.1.1. *The discretization $\mathscr{D}$ is called consistent on problem $\mathscr{P}$ if there exists a set $F^* \subset \mathrm{dom}(F)$ whose image $F(F^*)$ is dense in some neighbourhood of the point $0 \in Y$ and it is consistent at each element $v \in F^*$.*

The order of consistency in F^* is defined as $\inf\{p_v : v \in F^*\}$, where p_v denotes the order of consistency at point v.

Example 4.1.1. Let us consider the Examples 1.1.1, 1.1.2 and 1.1.3. Let us modify properly the operators F_n and ϕ_n in order to apply the explicit Euler method to the Cauchy problem (1.2)-(1.3). In sense of Definition 4.1.1 we verify consistency and its order on $F^* \subset \mathrm{dom}(F)$, where the sets are $\mathrm{dom}(F) := C^1([0,T])$ and $F^* := C^2([0,T])$. Then for the local discretization error we obtain

$$[F_n(\varphi_n(v)) - \psi_n(F(v))](t_k) = \begin{cases} \dfrac{v''(\theta_k)}{2K} & k = 1, \dots, K, \\ 0, & k = 0, \end{cases} \tag{4.1}$$

where $\theta_k \in (t_{k-1}, t_k)$ are given. Then $\|l_n(v)\|_{X_n} = \mathcal{O}(n^{-1})$ from Definition 1.2.3. Hence, for the class of problems with Lipschitz continuous right-hand side f, the explicit Euler method is consistent and the order of consistency equals one. ♣

As we have seen in Example A.1.1 the pointwise consistency at the solution does not imply convergence. One may think that the stronger consistency Definition 4.1.1 now ensures convergence. Example A.3.1 shows that this is not true.

Besides Assumptions 1.1.1 and 3.2.1 we assume the validity of the following assumptions. The first one is natural due to Remark 1.2.5.

Assumption 4.1.1. For problem $\mathscr{P}$ we assume that F^{-1} is continuous at the point $0 \in Y$.

The other ones related to the mappings φ_n and ψ_n.

Assumption 4.1.2. Let us apply the discretization $\mathscr{D}$ to problem $\mathscr{P}$. We assume that discretization $\mathscr{D}$ possesses the property: there exists $K_1 > 0$ such that for all $v \in \mathrm{dom}(F)$ the relation

$$\|\varphi_n(u^*) - \varphi_n(v)\|_{X_n} \leq K_1 \|u^* - v\|_X$$

holds for all $n \in \mathbb{I}$.

Assumption 4.1.3. We assume that discretization $\mathscr{D}$ possesses the property: there exists $K_2 > 0$ such that for all $y \in Y$ the relation

$$\|\psi_n(y) - \psi_n(0)\|_{Y_n} \leq K_2 \|y - 0\|_Y$$

holds for all $n \in \mathbb{I}$.

For the simplicity of the formulation, the collection of the Assumptions 1.1.1, 3.2.1 and 4.1.1-4.1.3 will be called Assumption $A^\star$.

Lemma 4.1.1. *Besides Assumption $A^\star$ we assume that*

i, discretization $\mathscr{D}$ on problem $\mathscr{P}$ is consistent,

ii, discretization $\mathscr{D}$ on problem $\mathscr{P}$ on the element u^* is K-stable with stability threshold R and constant C.

Then F_n is invertible at the point $\psi_n(0)$, i.e. there exists $F_n^{-1}(\psi_n(0))$ for sufficiently large indices n.

Proof. Due to the continuity of F^{-1} at the point $0 \in Y$ we can choose a sequence $(y^k)_{k\in\mathbb{I}}$ such that $y^k \to 0 \in Y$ and $F^{-1}(y^k) =: u^k \to u^*$. It follows that for some sufficiently large indices k the discretization $\mathscr{D}$ on problem $\mathscr{P}$ on the element u^k is K-stable with stability threshold $R/2$ and constant C. Moreover, F_n is continuous on $B_{R/2}(\varphi_n(u^k))$. Thus, for these indices k and also for sufficiently large n there exists $F_n^{-1} : B_{R/2C}(F_n(\varphi_n(u^k))) \to B_{R/2}(\varphi_n(u^k))$. According to Lemma 3.2.3, it is Lipschitz continuous with constant C. Let us write a trivial upper estimate

$$\left\|F_n(\varphi_n(u^k))\right\|_{Y_n} \leq \left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{Y_n} + \left\|\psi_n(F(u^k))\right\|_{Y_n}.$$

Due to consistency, the first term tends to 0 as $n \to \infty$. For the second term, based on Assumption 4.1.3 we have the estimate $\left\|\psi_n(y^k)\right\|_{Y_n} \leq K_2 \left\|y^k\right\|_{X_n}$. Since the right-hand side tends to 0 as $k \to \infty$, this means that the centre of the ball $B_{R/2C}(F_n(\varphi_n(u^k)))$ tends to $0 \in \mathcal{Y}_n$, which proves the statement. ∎

Corollary 4.1.2. *Under the conditions of Lemma 4.1.1, for sufficiently large indices k and n the following results are true.*

i, There exists $F_n^{-1}(\psi_n(y^k))$, since $\psi_n(y^k) \in B_{R/2C}(F_n(\varphi_n(u^k)))$.

ii, $F_n^{-1}(\psi_n(y^k)), \varphi_n(F^{-1}(y^k)) \in B_{R/2}(\varphi_n(u^*))$.

Analogously to the consistency, stability can also be defined on a set of points. This makes it possible to avoid the direct knowledge of the usually unknown u^*.

Definition 4.1.2. *The discretization $\mathscr{D}$ is called K-stable on problem $\mathscr{P}$ if there exist a positive constant C, $R \in (0, \infty]$ and a set $F^* \subset \mathrm{dom}(F)$ such that $u^* \in F^*$ and it is K-stable at each element $v \in F^*$ with stability threshold R and constant C.*

We reformulate our basic result, in which the notion of convergence is ensured by the notions of consistency and stability on a set.

Theorem 4.1.3. *Besides the Assumption $A^\star$ we suppose that discretization $\mathscr{D}$ on problem $\mathscr{P}$ is*

i, consistent,

ii, K-stable with some stability threshold R and constant C, respectively.

Then discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$ on the corresponding set F^*.

Proof. Using the triangle inequality, we have

$$\begin{aligned} \|\varphi_n(u^*) - u_n^*\|_{X_n} &= \left\|\varphi_n(F^{-1}(0)) - F_n^{-1}(\psi_n(0))\right\|_{X_n} \\ &\le \underbrace{\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{X_n}}_{(I)} \\ &+ \underbrace{\left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{X_n}}_{(II)} \\ &+ \underbrace{\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{X_n}}_{(III)}, \end{aligned} \tag{4.2}$$

where the elements $y^k \in Y$ are defined in the proof of Lemma 4.1.1.

In the next step we estimate the different terms on the right-hand side of (4.2).

(I) Based on Assumption 4.1.2 we have the estimate

$$\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{X_n} \le K_1 \left\|F^{-1}(0) - F^{-1}(y^k)\right\|_X .$$

Since $y^k \to 0$ as $k \to \infty$ and F^{-1} is continuous at the point $0 \in \mathcal{Y}$, therefore this term tends to 0 independently of n.

(II) This term can be written as $\left\|F_n^{-1}(F_n(\varphi_n(F^{-1}(y^k)))) - F_n^{-1}(\psi_n(y^k))\right\|_{X_n}$. Due to Corollary 4.1.2 we can use the stability estimate, therefore for this term we have

$$\begin{aligned} \left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{X_n} &\le C \left\|F_n(\varphi_n(F^{-1}(y^k))) - \psi_n(y^k)\right\|_{Y_n} \\ &= C \left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{Y_n} . \end{aligned}$$

Due to the consistency at element u^k the term on the right-hand side tends to 0.

(III) Due to Lemma 4.1.1 and Corollary 4.1.2 we can use the Lipschitz continuity of F_n^{-1} for the estimation of the third term. Hence, by using the Assumption 4.1.3, we have

$$\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{X_n} \leq C \left\|\psi_n(y^k) - \psi_n(0)\right\|_{Y_n} \leq CK_2 \left\|y^k\right\|_Y .$$

The right-hand side of the above estimate tends to 0 independently of the index n.

These estimations complete the proof. ■

Remark 4.1.1. In sense of Examples 1.1.1-1.1.3, Definitions 4.1.1, 4.1.2 and Theorem 4.1.3 we could easily analyze the stability and convergence property of the explicit Euler method for the problem (1.2)-(1.3). For the details see Example 40 in [27]. It is important to note that, as opposed to the usual direct proof of the convergence of the explicit Euler method, the convergence in this example yields the convergence on the whole space-time domain and not only at some fixed time level.

4.2 Relation between the basic notions

Under the Assumption $A^\star$ Theorem 4.1.3 shows us that, the consistency and stability of discretization $\mathscr{D}$ on problem $\mathscr{P}$ together imply the convergence, i.e. consistency and stability together form a sufficient condition for convergence. Obviously from this observation we cannot get an answer to the question of the necessity of these conditions.

However, one might ask that what is the general relation between the above listed notions. Since each of them can be true (T) or false (F), we have to consider eight different cases, listed in Table 4.1.

	Consistency	Stability	Convergence
1	T	T	T
2	T	T	F
3	T	F	T
4	T	F	F
5	F	T	T
6	F	T	F
7	F	F	T
8	F	F	F

Table 4.1. *The list of the different cases.*

We would like to note that Cases 6 and 8 in Table 4.1 are uninteresting from a practical point of view, therefore we neglected their investigation.

Based on this section we can theoretically answer the most important cases. Namely, these are Cases 1 and 2. Due to Examples A.3.2-A.3.4 taken from [27] we

can answer the basic question posed at beginning of this section. Using the numeration of the different cases in Table 4.1, the answers are included in Table 4.2. The results particularly show that neither consistency nor stability is a necessary condition for convergence.

Case	Answer	Reason
1	Always True	Theorem 4.1.3
2	Always False	Theorem 4.1.3
3	Possible	Example A.3.2
4	Possible	Examples A.1.1 and A.3.1
5	Possible	Example A.3.3
6	n.a.	n.a.
7	Possible	Example A.3.4
8	n.a.	n.a.

Table 4.2. *The answers of the posed question.*

CHAPTER 5

Results of the thesis

This dissertation deals with stability concepts for operator equations and their possible application areas in theoretical numerical analysis.

Chapter 1

In Chapter 1 we are interested in how we can define the basic notions for nonlinear operator equations. Our framework is inspired by Stetter's framework and the papers of Sanz-Serna, Palencia and López-Marcos, who systemically studied basic questions in this area.

Sections 1.1 and 1.2 are based on the Author's paper [27]. Our framework is a modified version of Stetter's framework.

In Section 1.1 we set the problem with the help of the Definitions 1.1.1, 1.1.2 and 1.1.3 between the continuous problem (1.1) and the discrete problem (1.4). Examples 1.1.1, 1.1.2 and 1.1.3 help to understand the introduced framework through the initial-value problem (1.2)-(1.3) with the applied explicit Euler numerical method.

In Section 1.2 we define the basic notions: convergence and consistency. Having defined convergence in Definition 1.2.2 in Remark 1.2.2 we mentioned the other approach. The consistency Definition 1.2.4 helps us getting information about the behaviour of the global discretization error. Remark 1.2.5 points out that consistency in itself does not imply convergence, therefore we need an additional condition.

Chapter 2

The introductory part of Chapter 2 motivates the notion of N-stability and introduces it in Definition 2.0.5, which was originally defined by López-Marcos and Sanz-Serna in [44]. Based on the Author's paper [29] in Theorem 2.0.1 we show that this notion fulfills the basic theorem of numerical analysis for the nonlinear case.

However, the main goal of this chapter is to show the benefits of N-stability in theoretical numerical analysis. These are the following.

- ◇ Linear stability is a special case (Section 2.1)
- ◇ Zero-stability and operator form of multistep methods (Section 2.2)

◇ New stability technique for time-dependent problems (Section 2.3)

◇ Numerical stability for nonlinear abstract Cauchy problems (Section 2.4)

Results of Chapter 2 are based on the Author's papers. Namely, the results of Sections 2.1, 2.2, 2.3 and 2.4 can be found in paper [28], preprint [25], papers [29], [28] and the accepted paper [19], respectively.

In Section 2.1 we deal with the linear version of (1.4). From the investigation it turns out that N-stability can be viewed as the natural extension of the classical linear stability definition for the linear problems. Due to Remark 2.1.1 the bound (2.4) in Definition 2.1.1 implies the existence and uniqueness of the solutions of the linear problems, the uniform boundedness of the inverse operator and the basic theorem of numerical analysis. In the end of this section we show that for linear problems N-stability is equivalent to the classical linear stability notion.

The main idea behind Section 2.2 is to apply our framework and the N-stability notion to prove zero-stability of multistep methods. We show how the scalar initial-value problem (2.5)-(2.6) fits into our framework. As an application of N-stability, in Theorem 2.0.1 we prove the well-known zero-stability theorem for one-step methods. Table 2.1 summarizes the choices of operators, normed spaces and corresponding norms in order to prove the above mentioned theorem. In Table 2.2 we sum up how the one-step zero-stability definitions from the literature fits into out framework. In the next train of thought we extend this approach for multistep methods.

Since a lot of physical, biological or chemical processes can be fit into our framework, in Section 2.3 we deal with two classical problem classes: reaction-diffusion problems and advection problems. Considering these benchmark problems our goal is to show that N-stability notion can serve as a new and effective tool for verifying the stability of a given method for time-dependent problems. In case of the diffusion problem (2.18)-(2.20) and the reaction-diffusion problem (2.32)-(2.34) we verify the N-stability of an IMEX-method (θ-method in time) in the introduced norms. These results correspond to Theorem 2.3.1 and 2.3.3 respectively. Table 2.5 reviews that with this N-stability approach we get back the well-known convergence results of the literature. Similarly, using this technique for advection problems (2.40)-(2.42) and (2.54)-(2.56) Theorems 2.3.4 and 2.3.6 prove the N-stability of the centralized Crank–Nicolson IMEX-method in the introduced norm. Table 2.6 shows the appropriate choices to prove the above mentioned theorems in these benchmark problems.

In Section 2.4 we demonstrate the application of the N-stability notion for one-parameter semigroups for linear and nonlinear evolution equations. In the first part of this section we briefly summarize the general Lax equivalence theorem for linear operator equations, which was proven by Palencia and Sanz-Serna. Furthermore, as an application of their theorem we give two examples (the well-posed homogeneous abstract Cauchy problem and the well-posed inhomogeneous abstract Cauchy problem in L_2) in which we show how we can get back from this theorem the semigroup case. In the end of the first part we mentioned the most used numerical techniques for operator semigroups. These motivate the use of our framework, the N-stability notion and the rational-type temporal discretizations in order to prove numerical stability of nonlinear abstract Cauchy-problems. Using

the fundamental results of Brezis, Crandall, Liggett and Pazy in this field we prove in Theorem 2.4.6 that in case of ω-dissipative operators nonlinear rational-type temporal discretizations are N-stable. From the numerical analysis perspective we extend the applicable class of numerical methods for the nonlinear abstract Cauchy problems with ω-dissipative operators. In the end of this part we apply our result to the linear case and we show that it coincides with the result of Palencia and Sanz-Serna.

Chapter 3

Chapter 3 deals with other stability notions for operator equations. In this chapter our main goal is to use more sophisticated stability notions and prove corresponding theoretical results. Furthermore, we would like to comment shortly on other existing stability notions. Short outline of Chapter 3:

- ◇ Incompleteness of N-stability (Section 3.1)
- ◇ K-stability and theoretical results (Section 3.2)
- ◇ T-stability and theoretical results (Section 3.3)
- ◇ Notes on further stability notions (Section 3.4)

Results of Chapter 3 are based on the Author's papers. Namely, the results of Sections 3.1, 3.2 and 3.3 can be found in paper [28] and [24], respectively.

Taking a simple Ricatti-type equation we show the necessity of the N-stability notion in Section 3.1. Namely, in this example in case of explicit Euler method the N-stability definition is too restrictive, since an arbitrary chosen element is too far from its perturbation. This motivates to introduce the idea of local stability and stability threshold notions.

In Section 3.2 we make the first step towards this direction using the semistability Definition 3.2.1. After the K-stability Definition 3.2.2 we give theoretical results. Based on a lemma of Stetter we prove Theorem 3.2.5. This theorem guarantees that (1.4) has a unique solution in some suitably chosen ball. It means that in the nonlinear case K-stability locally satisfies the properties that the linear stability notion (or, equivalently, the N-stability notion for the linear case) does. In the second part of this section we prove K-stability of the explicit Euler method for a general class of operators in Theorems 3.2.6 and 3.2.8. Due to these results we simultaneously prove the K-stability of the explicit Euler method for the Ricatti-type problem of Section 3.1.

In Section 3.3 we introduce a completely different approach to define nonlinear stability. It was originally defined by Trenogin and in Definition 3.3.1 we called it T-stability. In Theorem 3.3.2 we prove that in case of T-stability the basic theory of numerical analysis holds for the nonlinear case. However, in this section our main goal is to improve Trenogin's original result. He proved that the explicit Euler method is T-stable for the initial-value problem (1.2)-(1.3) on an equidistant grid. In contrast with Trenogin we prove that an arbitrary one-step method is T-stable both on the equidistant and non-equidistant grids. The corresponding T-stability constants are summarized in Table 3.2.

In Section 3.4 we give a brief summary of our thoughts about the so-called S-stabiliy and LSS-stability notions.

Chapter 4

Results of Chapter 4 are based on the Author's paper [27].

In the first part of Section 4.1 we extend the previously given pointwise (local) definitions to the set (global) ones. The reason behind this idea is that in some sense our strongest result Theorem 3.2.5 requires to check the K-stability and the consistency on the unknown solution of (1.1). Under reasonable assumptions we prove the set version of the basic theorem of numerical analysis. In the second part we show the relation between the basic notions. Since consistency, stability and convergence can be true or false, we consider eight different cases, which are listed in Table 4.1. We neglect two cases, since these are uninteresting from a practical point of view. Based on the previous results of this section we can theoretically answer the most important cases and we can also give examples in the Appendix Section A.3. Using the numeration of the different cases in Table 4.1, the answers are included in Table 4.2.

APPENDIX A

Appendix related to the Chapters

A.1 Chapter 1

Example A.1.1. Let us consider the case

$$X = X_n = Y = Y_n = \mathbb{R},$$

$$\mathrm{dom}(F) = \mathrm{dom}(F_n) = [0, \infty),$$

$$\varphi_n = \psi_n = \text{Identity}.$$

Our aim is to solve the scalar equation $F(u) = 0$, where we assume that it has a unique solution $u^* = 0$. We define the F_n operator as

$$F_n(v) = \frac{1 - v}{n}, \quad v \in X,\ n \in \mathbb{I}.$$

Due to the linearity of the mappings φ_n and ψ_n, we have $l_n = F_n(0) - 0 = F_n(0)$. The discretization is consistent, since $F_n(0) \to 0$ if n tends to ∞. However, it is not convergent, since the solution of the discrete problems is $u_n^* = 1$. ♣

A.2 Chapter 2

Lemma A.2.1. *The operator* (2.7) *is injective.*

Proof. Injectivity of operator (2.7) means that if $w_1(t), w_2(t) \in C^1([0, T]$ such that $[Lw_1](t) = [Lw_2](t)$, then $w_1(t) = w_2(t)$. Due to the form $Lu = g$ we have

$$w_1'(t) - f(t, w_1(t)) = w_2'(t) - f(t, w_2(t)),\ t \in (0, T]$$

and

$$w_1(0) = w_2(0).$$

We introduce the function $r(t) = w_1'(t) - f(t, w_1(t)) = w_2'(t) - f(t, w_2(t))$. Then the function r is a given continuous function. Obviously the initial-value problem

$$\begin{aligned} w'(t) &= f(t, w) + r(t), \quad && t \in (0, T], \\ w(0) &= \text{given}, && t = 0 \end{aligned}$$

has a unique solution, since the right-hand side function is a Lipschitz continuous function with respect to its second variable and f has the same Lipschitz constant. Due to the definition of r we have

$$w_1'(t) = f(t, w_1(t)) + r(t),\ \ w_2'(t) = f(t, w_2(t)) + r(t) \text{ and } w_1(0) = w_2(0).$$

Due to the uniqueness of the previously showed initial-value problem it follows that $w_1(t) = w_2(t)$.

■

Lemma A.2.2. *The operator* (2.9) *is injective.*

Proof. Injectivity of the operator (2.9) means that if $z_n, w_n \in \mathbb{F}(\omega_\tau)$ such that $L_n z_n = L_n w_n$, then $z_n = w_n$. Taking into account the definition of operator (2.9) we have $z_n(0) = w_n(0)$. Since

$$\Phi(\tau_1, t_0, z_n(t_0), z_n(t_1)) = \Phi(\tau_1, t_0, w_n(t_0), w_n(t_1)).$$

The common part is denoted by r_1. The unknown $z_n(t_1)$ and $w_n(t_1)$ are uniquely determined from

$$\Phi(\tau_1, t_0, z_n(t_0), z_n(t_1)) = r_1(t) \text{ and } \Phi(\tau_1, t_0, w_n(t_0), w_n(t_1)) = r_1(t).$$

Then $z_n(t_1) = w_n(t_1)$. Applying this process in the previous train of thought one can conclude that $z_n = w_n$. ■

Lemma A.2.3. *The following relation holds:*

$$\left\|Q_1^{-1}Q_2\right\|_2 = 1. \tag{A.1}$$

Proof. The matrix D_p in (2.47) is a skew-symmetric matrix ($D_p^* = -D_p$). Moreover, for an arbitrary matrix $M \in \mathbb{R}^{n\times n}$ we have the relation $\|M\|_2^2 = \rho(MM^*)$. Using these properties to (A.1), we obtain

$$\begin{aligned}
\left\|Q_1^{-1}Q_2\right\|_2^2 &= \left\|(I+D_p)^{-1}(I-D_p)\right\|_2^2 \\
&= \rho\left((I+D_p)^{-1}(I-D_p)\Big[(I+D_p)^{-1}(I-D_p)\Big]^*\right) \\
&= \rho\left((I+D_p)^{-1}(I-D_p)(I-D_p)^*\Big[(I+D_p)^{-1}\Big]^*\right) \\
&= \rho\left((I+D_p)^{-1}(I-D_p)(I+D_p)\Big[(I+D_p)^{-1}\Big]^*\right) \\
&= \rho\left((I+D_p)^{-1}(I+D_p)(I-D_p)\Big[(I+D_p)^{-1}\Big]^*\right) \\
&= \rho\left((I-D_p)\Big[(I+D_p)^{-1}\Big]^*\right) = \rho\left((I+D_p)^{-1}(I-D_p)^*\right) \\
&= \rho\left((I+D_p)^{-1}(I+D_p)\right) = 1.
\end{aligned}$$

This relation proves our statement. ■

Lemma A.2.4. *Let us consider the Cauchy problem (1.2)-(1.3). Then for the problem (1.2)-(1.3) the local discretization error of the forward Euler method on an equidistant grid can be estimated by*

$$l_n(u^*)(t_k) \leq \frac{M_2(u^*)}{2}h,$$

where $t_k = kh,\ k = 0, \ldots, K,\ M_2(u^*) := \sup_{t\in[0,1]} |(u^*)''(t)| < \infty$ *and* h *is the step-size of the grid.*

Proof. We have the relation

$$\begin{aligned} l_n(u^*)(t_k) = [F_n(\varphi_n(u^*))](t_k) &= \frac{u^*(t_k) - u^*(t_{k-1})}{h} - (u^*)'(t_{k-1}) \\ &\leq \max_{0\leq k\leq K} \left|(u^*)'((k-1)h) - \frac{1}{h}\Big(u^*(kh) - u^*((k-1)h)\Big)\right| \\ &= \max_{0\leq k\leq K} \left|\frac{1}{h}\int_{(k-1)h}^{kh} (u^*)'((k-1)h) - (u^*)'(s)\mathrm{ds}\right| \\ &\leq \frac{1}{h}\max_{0\leq k\leq K} \int_{t_{k-1}}^{t_k} |(u^*)'(t_{k-1}) - (u^*)'(s)|\mathrm{ds}. \end{aligned}$$

Hence,

$$l_n(u^*)(t_k) \leq \frac{1}{h}M_2(u^*)\frac{1}{2}h^2 = \frac{M_2(u^*)}{2}h.$$

■

A.3 Chapter 4

In this section the following examples correspond to Cases 3,4,5 and 7 of Table 4.2. These examples are taken from [27].

Example A.3.1. Let us consider the case

$$X = X_n = Y = Y_n = \mathbb{R},$$

$$\varphi_n = \psi_n = \text{Identity}.$$

We would like to solve the scalar $F(u) = 0$, where the function $F \in C(\mathbb{R}, \mathbb{R})$ is given as

$$F(x) = \begin{cases} |x|\,, & \text{if } x \in (-1, 1)\,, \\ 1, & \text{if } x \in (-\infty, -1] \cup [1, \infty)\,. \end{cases}$$

Obviously this problem has a unique solution $u^* = 0$. We define the operator F_n as

$$F_n(x) = \begin{cases} \frac{1}{n}, & \text{if } x \in \left[-\frac{1}{n}, \frac{1}{n}\right], \\ x, & \text{if } x \in \left(\frac{1}{n}, 1\right), \\ 1, & \text{if } x \in (-\infty, -1] \cup [1, n) \cup [n+2, \infty), \\ -x, & \text{if } x \in \left(-1, -\frac{1}{n}\right), \\ |x-(n+1)|, & \text{if } x \in [n, n+2). \end{cases}$$

This discretization is consistent on the entire $\mathbb{R}$. However, it is not convergent, since the solution of the discrete problems is $u_n^* = n + 1$. ♣

In the following three examples the normed spaces and the corresponding mappings will be the same. Namely,

$$X = X_n = Y = Y_n = \mathbb{R},$$

$$\mathrm{dom}(F) = \mathrm{dom}(F_n) = [0, \infty),$$

$$\varphi_n = \psi_n = \text{identity}.$$

Our aim is to solve the scalar equation

$$F(v) = v^2 = 0, \tag{A.2}$$

which has the unique solution $u^* = 0$.

Example A.3.2. In order to solve equation (A.2) we choose the numerical method defined by the nth Lagrangian interpolation, i.e. $F_n(v)$ is the Lagrangian interpolation polynomial of order n.

Since the Lagrange interpolation is exact for $n \geq 2$, therefore $F_n(v) = v^2$ holds for all $n \geq 2$. Hence, clearly the numerical method is consistent and convergent. The operator F_n^{-1} can be defined easily and it is $F_n^{-1}(v) = \sqrt{v}$. However, its derivative is not bounded around the point $u^* = 0$, therefore the numerical method is not stable. ♣

Example A.3.3. For solving equation (A.2) we choose now the numerical method as $F_n(v) = 1 - nv$. The roots of the discrete equations $F_n(v) = 0$ are $u_n^* = 1/n$, therefore $u_n^* \to u^* = 0$ as $n \to \infty$. This means that the numerical method is convergent. We observe that $\varphi_n(F_n(0)) = \varphi_n(1) = 1$ and $\psi_n(F(0)) = \psi_n(0) = 0$. Hence, for the local discretization error we have $|l_n| = 1$ for any indices n. This means that the numerical method is not consistent.

One can easily check that F_n is invertible and $F_n^{-1}(v) = -v/n + 1/n$. Thus, the derivative of the inverse operators are uniformly bounded on $[0, \infty)$ by 1 for any n. Therefore the numerical method is stable. ♣

Example A.3.4. We would like to solve equation (A.2). Choosing the numerical method $F_n(x) = 1 - nx^2$ we can conclude that $u_n^* = 1/\sqrt{n}$. Therefore, $u_n^* \to u^* = 0$ as $n \to \infty$. This means that the numerical method is convergent.

However, due to the relations $\varphi_n(F_n(0)) = \varphi_n(1) = 1$ and $\psi_n(F(0)) = \psi_n(0) = 0$ the defined method is not consistent. It is not stable, since the inverse can be written as $F_n^{-1}(v) = \sqrt{(1-v)/n}$, i.e. derivatives are not bounded. ♣

Bibliography

[1] L. M. Abia, O. Angulo, J. C. López-Marcos, and M. A. López-Marcos, *Numerical integration of a hierarchically size-structured population model with contest competition*, J. Comput. Appl. Math., 258 (2014), pp. 116–134.

[2] O. Angulo, J. C. López-Marcos, and M. A. López-Marcos, *Analysis of an efficient integrator for a size-structured population model with a dynamical resource*, Comput. Math. Appl., 68 (2014), pp. 941–961.

[3] U. M. Ascher, *Numerical methods for evolutionary differential equations*, vol. 5 of Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[4] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167. Special issue on time integration (Amsterdam, 1996).

[5] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[6] K. A. Bagrinovskii and S. K. Godunov, *Difference schemes for multidimensional problems*, Dokl. Akad. Nauk SSSR (N.S.), 115 (1957), pp. 431–433.

[7] A. Bátkai, P. Csomós, B. Farkas, and G. Nickel, *Operator splitting with spatial-temporal discretization*, in Spectral theory, mathematical system theory, evolution equations, differential and difference equations, vol. 221 of Oper. Theory Adv. Appl., Birkhäuser/Springer Basel AG, Basel, 2012, pp. 161–171.

[8] A. Bátkai, P. Csomós, and G. Nickel, *Operator splittings and spatial approximations for evolution equations*, J. Evol. Equ., 9 (2009), pp. 613–636.

[9] A. Belleni-Morante and A. C. McBride, *Applied nonlinear semigroups*, vol. 3 of Wiley Series in Mathematical Methods in Practice, John Wiley & Sons, Ltd., Chichester, 1998. An introduction.

[10] P. Brenner and V. Thomée, *On rational approximations of semigroups*, SIAM J. Numer. Anal., 16 (1979), pp. 683–694.

[11] H. Brézis and A. Pazy, *Convergence and approximation of semigroups of nonlinear operators in Banach spaces*, J. Functional Analysis, 9 (1972), pp. 63–74.

[12] L. E. J. Brouwer, *Beweis der Invarianz der geschlossenen Kurve*, Math. Ann., 72 (1912), pp. 422–425.

[13] F. E. Browder, *Nonlinear equations of evolution and nonlinear accretive operators in Banach spaces*, Bull. Amer. Math. Soc., 73 (1967), pp. 867–874.

[14] J. C. Butcher, *Numerical methods for ordinary differential equations*, John Wiley & Sons, Ltd., Chichester, second ed., 2008.

[15] J. G. Charney, R. Fjörtoft, and J. von Neumann, *Numerical integration of the barotropic vorticity equation*, Tellus, 2 (1950), pp. 237–254.

[16] R. Courant, K. Friedrichs, and H. Lewy, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann., 100 (1928), pp. 32–74.

[17] M. G. Crandall and T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.

[18] M. G. Crandall and A. Pazy, *Nonlinear evolution equations in Banach spaces*, Israel J. Math., 11 (1972), pp. 57–94.

[19] P. Csomós, I. Faragó, and I. Fekete, *Numerical stability for nonlinear evolution equations*, Comput. Math. Appl., accepted (2015).

[20] P. Csomós, I. Faragó, and Á. Havasi, *Weighted sequential splittings and their analysis*, Comput. Math. Appl., 50 (2005), pp. 1017–1031.

[21] K.-J. Engel and R. Nagel, *One-parameter semigroups for linear evolution equations*, vol. 194 of Graduate Texts in Mathematics, Springer-Verlag, New York, 2000. With contributions by S. Brendle, M. Campiti, T. Hahn, G. Metafune, G. Nickel, D. Pallara, C. Perazzoli, A. Rhandi, S. Romanelli and R. Schnaubelt.

[22] ——, *A short course on operator semigroups*, Universitext, Springer, New York, 2006.

[23] I. Faragó, *Convergence and stability constant of the theta-method*, in Applications of mathematics 2013, Acad. Sci. Czech Repub. Inst. Math., Prague, 2013, pp. 42–51.

[24] I. Faragó and I. Fekete, *T-stability of general one-step methods for abstract initial-value problems*, Open Math. J., 6 (2013), pp. 19–25.

[25] ——, *0-stability of operator form of linear multistep methods*, preprint (2015).

[26] I. Faragó and J. Geiser, *Iterative operator-splitting methods for linear problems*, International Journal of Computational Science and Engineering, 3 (2007), pp. 255–263.

[27] I. Faragó, M. E. Mincsovics, and I. Fekete, *Notes on the basic notions in nonlinear numerical analysis*, in The 9th Colloquium on the Qualitative Theory of Differential Equations, vol. 6 of Proc. Colloq. Qual. Theory Differ. Equ., Electron. J. Qual. Theory Differ. Equ., Szeged, 2012, pp. 1–22.

[28] I. Fekete and I. Faragó, *N-stability of the θ-method for reaction-diffusion problems*, Miskolc Math. Notes, 15 (2014), pp. 447–458.

[29] ——, *Stability concepts and their applications*, Comput. Math. Appl., 67 (2014), pp. 2158–2170.

[30] W. Gautschi, *Numerical analysis*, Birkhäuser Basel, 2nd ed., 2012. An introduction.

[31] Gavurin, M. K., *Lectures on Computational Methods (in Russian)*, Izdat. Nauka, Moscow, 1971.

[32] J. A. Goldstein, *Approximation of nonlinear semigroups and evolution equations*, J. Math. Soc. Japan, 24 (1972), pp. 558–573.

[33] W. Hackbusch, *The concept of stability in numerical mathematics*, vol. 45 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2014.

[34] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations. I*, vol. 8 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, second ed., 1993. Nonstiff problems.

[35] E. Hairer and G. Wanner, *Solving ordinary differential equations. II*, vol. 14 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2010. Stiff and differential-algebraic problems, Second revised edition, paperback.

[36] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numer., 19 (2010), pp. 209–286.

[37] K. Ito and F. Kappel, *Evolution equations and approximations*, vol. 61 of Series on Advances in Mathematics for Applied Sciences, World Scientific Publishing Co., Inc., River Edge, NJ, 2002.

[38] L. V. Kantorovich, *Functional analysis and applied mathematics*, Uspehi Matem. Nauk (N.S.), 3 (1948), pp. 89–185.

[39] T. Kato, *Trotter's product formula for an arbitrary pair of self-adjoint contraction semigroups*, in Topics in functional analysis (essays dedicated to M. G. Kreuın on the occasion of his 70th birthday), vol. 3 of Adv. in Math. Suppl. Stud., Academic Press, New York-London, 1978, pp. 185–195.

[40] H. B. Keller, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29 (1975), pp. 464–474.

[41] J. D. Lambert, *Numerical methods for ordinary differential systems*, John Wiley & Sons, Ltd., Chichester, 1991. The initial value problem.

[42] P. D. Lax and R. D. Richtmyer, *Survey of the stability of linear finite difference equations*, Comm. Pure Appl. Math., 9 (1956), pp. 267–293.

[43] R. J. LeVeque, *Finite difference methods for ordinary and partial differential equations*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007. Steady-state and time-dependent problems.

[44] J. C. López Marcos and J. M. Sanz-Serna, *A definition of stability for nonlinear problems*, in Numerical treatment of differential equations (Halle, 1987), vol. 104 of Teubner-Texte Math., Teubner, Leipzig, 1988, pp. 216–226.

[45] ——, *Stability and convergence in numerical analysis. III. Linear investigation of nonlinear stability*, IMA J. Numer. Anal., 8 (1988), pp. 71–84.

[46] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.

[47] J. K. Mountain, *The Lax equivalence theorem for linear, inhomogeneous equations in L^2 spaces*, J. Approx. Theory, 33 (1981), pp. 126–130.

[48] G. Nickel, *Evolution semigroups and product formulas for nonautonomous Cauchy problems*, Math. Nachr., 212 (2000), pp. 101–116.

[49] T. Ortega and J. M. Sanz-Serna, *Nonlinear stability and convergence of finite-difference methods for the "good" Boussinesq equation*, Numer. Math., 58 (1990), pp. 215–229.

[50] C. Palencia and J. M. Sanz-Serna, *Equivalence theorems for incomplete spaces: an appraisal*, IMA J. Numer. Anal., 4 (1984), pp. 109–115.

[51] ——, *An extension of the Lax-Richtmyer theory*, Numer. Math., 44 (1984), pp. 279–283.

[52] A. Pazy, *Semigroups of linear operators and applications to partial differential equations*, vol. 44 of Applied Mathematical Sciences, Springer-Verlag, New York, 1983.

[53] J. M. Sanz-Serna, *Stability and convergence in numerical analysis. I. Linear problems—a simple, comprehensive account*, in Nonlinear differential equations (Granada, 1984), vol. 132 of Res. Notes in Math., Pitman, Boston, MA, 1985, pp. 64–113.

[54] J. M. Sanz-Serna and C. Palencia, *A general equivalence theorem in the theory of discretization methods*, Math. Comp., 45 (1985), pp. 143–152.

[55] J. M. Sanz-Serna and J. G. Verwer, *A study of the recursion $y_{n+1} = y_n + \tau y_n^m$*, J. Math. Anal. Appl., 116 (1986), pp. 456–464.

[56] H. J. Stetter, *Analysis of discretization methods for ordinary differential equations*, Springer-Verlag, New York-Heidelberg, 1973. Springer Tracts in Natural Philosophy, Vol. 23.

[57] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[58] J. C. Strikwerda, *Finite difference schemes and partial differential equations*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 2004.

[59] E. Süli and D. F. Mayers, *An introduction to numerical analysis*, Cambridge University Press, Cambridge, 2003.

[60] J. W. Thomas, *Numerical partial differential equations: finite difference methods*, vol. 22 of Texts in Applied Mathematics, Springer-Verlag, New York, 1995.

[61] J. A. Trangenstein, *Numerical solution of hyperbolic partial differential equations*, Cambridge University Press, Cambridge, 2009.

[62] V. A. Trenogin, *Functional analysis (in Russian)*, Nauka, Moscow, 1980.

[63] H. F. Trotter, *Approximation of semi-groups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.

[64] G. Wanner, *Dahlquist's classical papers on stability theory*, BIT, 46 (2006), pp. 671–683.

Acknowledgements

I would like to thank my thesis advisor, Prof. István Faragó, for making me fascinated about numerical analysis during my undergraduate studies and for having faith in me. Working on my B.Sc. and M.Sc. theses under his supervision, I could delve into the world of numerical analysis. Having supervised my Ph.D., he had always turned my attention towards interesting problems and relations, which were invaluable for writing my thesis. I am proud to call him my mentor and even my friend, since I could discuss with him all the aspects of my life any time. For that I shall remain eternally grateful.

I am glad to have had the opportunity to work with Dr. Petra Csomós (MTA-ELTE Numerical Analysis and Large Networks Research Group) and thanks to her I gained a deeper understanding of numerical methods for operator semigroups.

I am grateful to the Professors, my fellow students and the administrators of the Department of Applied Analysis and Computational Mathematics, who made my time here unforgettable and enjoyable.

I would like to express my gratitude to Prof. David Ketcheson (KAUST), who raised the question of the topic of Section 2.2. I also thank Prof. Sanz-Serna (University Carlos III de Madrid), who supported me in that it is worth dealing with Section 2.4.

I am grateful for generous funding during my Ph.D. studies, provided by the MTA-ELTE Numerical Analysis and Large Networks Research Group. Some parts of Chapter 2 was supported by the European Union and the State of Hungary, co-financed by the European Social Fund witihin the framework of TAMOP-4.2.4.A/2-11/1-2012-0001 'National Program of Excellence'.

I would like to thank my family for giving me the ideal environment during my studies. I always think about them with love. I cannot find the words to express my gratitude towards my parents for sacrifising so much in order to help me study mathematics at Eötvös Loránd University. I dedicate this Ph.D. thesis to them.

Still, it takes more than numerical analysis to make life worth living. For this I thank my wife, who is as roguish as a cat.

Budapest, May 26, 2015 *Imre Fekete*

Summary

This dissertation deals with stability concepts for operator equations and their possible application areas in theoretical numerical analysis. This thesis is based on the Author's papers [27], [24], [29], [28], the accepted paper [19] and the preprint [25]. The thesis consists of five chapters.

In Chapter 1 we set the problem in an abstract setting and introduce the basic notions in numerical analysis. Furthermore, we show what is the relation between consistency and convergence for nonlinear operator equations.

In Chapter 2 we deal with N-stability notion and we show its possible application areas in theoretical numerical analysis. In Section 2.2 it turns out that linear multistep methods and the zero-stability notion fits into our framework and we regain the classical results from the literature. In Section 2.3 we offer a new and effective tool in order to verify stability results for time-dependent problems. The benchmark problems are reaction-diffusion and transport problems. In Section 2.4 we consider nonlinear evolution equations whose solution is given by a nonlinear semigroup. We show that the definition of nonlinear semigroups already contains a sort of time discretization, the implicit Euler method, which leads to N-stable discrete problems when applied together with certain convergent space discretizations. Moreover, we propose a more general time discretization, being the nonlinear counterpart of the rational approximations in the linear case and show its N-stability as well.

In Chapter 3 we deal with other stability notions. First, in Section 3.1 we give an example to motivate local type stability notions. In Section 3.2 we show the benefits of this notion in theory as well as from the application point of view. In Section 3.3 we prove theoretical results for Trenogin's stability notion and we improve his results. In the end of this chapter we give some comments on other stability notions.

In the fist part of Chapter 4 we extend the previously given pointwise (local) definitions to the set (global) ones. Under reasonable assumptions we prove the set version of the basic theorem of numerical analysis. In the second part we show the relation between the basic notions. Based on the previous results of this section we can theoretically answer the most important cases and we can also give examples in the Appendix Section A.3.

Összefoglalás

A disszertáció az operátoregyenletek stabilitási koncepciójával és azok elméleti numerikus analízisbeli alkalmazási lehetőségeivel foglalkozik. A dolgozat a Szerző megjelent cikkjein [27], [29], [28], [24], elfogadott cikkjén [19] és kéziratán [25] alapszik. A disszertáció öt fejezetből áll.

Az 1. Fejezetben absztrakt környezetben fogalmazzuk meg a problémát és definiáljuk a hozzá szükséges numerikus analízisbeli alapfogalmakat. Továbbá nemlineáris operátoregyenletek esetén megmutatjuk a kapcsolat a konzisztencia és a konvergencia között.

A 2. Fejezetben az N-stabilitás fogalmával és elméleti numerikus analízisbeli alkalmazási területeivel foglalkozunk. A 2.2. Fejezetből az derül ki, hogy a lineáris többlépéses módszerek, valamint a zéró-stabilitás illik az absztrakt környezetünkbe és ennek segítségével visszakapjuk az irodalomból ismert klasszikus eredményeket. A 2.3. Fejezetben egy új és hatékony technikát mutatunk időfüggő feladatok stabilitásvizsgálatához. Alapproblémának a reakció-diffúzió és transzport egyenleteket választjuk. A 2.4. Fejezetben nemlineáris evolúciós egyenleteket tekintünk, melyek megoldásai nemlineáris félcsoportot alkotnak. Megmutatjuk, hogy a nemlineáris operátorfélcsoportok definíciója is tartalmaz egyfajta időbeli diszkretizációt (implicit Euler), mely konvergens térbeli diszkretizációval együtt N-stabil diszkrét feladatok sorozatához vezet. Továbbá egy általános idődiszkretizációs módszert javaslunk, mely a racionális approximáció nemlineáris változatának tekinthető és megmutatjuk ennek az N-stabilitását is.

A 3. Fejezet további stabilitás fogalmakkal foglalkozik. Először a 3.1. Fejezetben motiváljuk a lokális típusú stabilitási fogalmakat. A 3.2. Fejzetben ennek mind elméleti mind alkalmazhatósági előnyeit is ismertetjük. A 3.3. Fejezetben Trenogin stabilitási elméletét alkalmazva további elméleti eredményeket bizonyítunk, illetve élesítjük korábbi eredményeit. A fejezetet további stabilitási fogalmakhoz kapcsolódó megjegyzéseinkkel zárjuk.

A 4. Fejezet első részében kiterjesztjük a korábbi elem alapú (lokális) definícióinkat halmaz (globális) alapúra. Értelmes feltevések mellett bizonyítjuk a halmaz alapú numerikus analízis alaptételét. A második részben megmutatjuk az alapfogalmak közötti kapcsolatot. A fejezet korábbi részében elért elméleti eredmények alapján elméleti úton válaszolunk a legfontosabb esetekre, emellett néhány további példát is mutatunk.

[38]**ADATLAP**

a doktori értekezés nyilvánosságra hozatalához

I. A doktori értekezés adatai

A szerző neve: Fekete Imre.....

MTMT-azonosító: 10033666.....

A doktori értekezés címe és alcíme: „Stability concepts and their applications"

DOI-azonosító[39]: 10.15476/ELTE.2015.093 .....

A doktori iskola neve: Matematika Doktori Iskola.....

A doktori iskolán belüli doktori program neve: Alkalmazott Matematika.....

A témavezető neve és tudományos fokozata: Dr. Faragó István, MTA Doktora

A témavezető munkahelye:

Alkalmazott Analízis és Számításmatematikai Tanszék, ELTE és

MTA-ELTE Numerikus Analízis és Nagy Hálózatok Kutatócsoport.....

II. Nyilatkozatok

A doktori értekezés szerzőjeként[40]

a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom a Természettudományi Kar Tudományszervezési és Egyetemközi Kapcsolatok Osztályának ügyintézőjét .....Bíró Évát....., hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[41]

c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés *(dátum)*-ig tartó időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;[42]

d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.[43]

2. A doktori értekezés szerzőjeként kijelentem, hogy

a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti , önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt.: Budapest, 2015. május 26.

.....Fekete Imre.....

a doktori értekezés szerzőjének aláírása

[38] Beiktatta az Egyetemi Doktori Szabályzat módosításáról szóló CXXXIX/2014. (VI. 30.) Szen. sz. határozat. Hatályos: 2014. VII.1. napjától.

[39] A kari hivatal ügyintézője tölti ki.

[40] A megfelelő szöveg aláhúzandó.

[41] A doktori értekezés benyújtásával egyidejűleg be kell adni a tudományági doktori tanácshoz a szabadalmi, illetőleg oltalmi bejelentést tanúsító okiratot és a nyilvánosságra hozatal elhalasztása iránti kérelmet.

[42] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a minősített adatra vonatkozó közokiratot.

[43] A doktori értekezés benyújtásával egyidejűleg be kell nyújtani a mű kiadásáról szóló kiadói szerződést.